MASTER'S THESIS

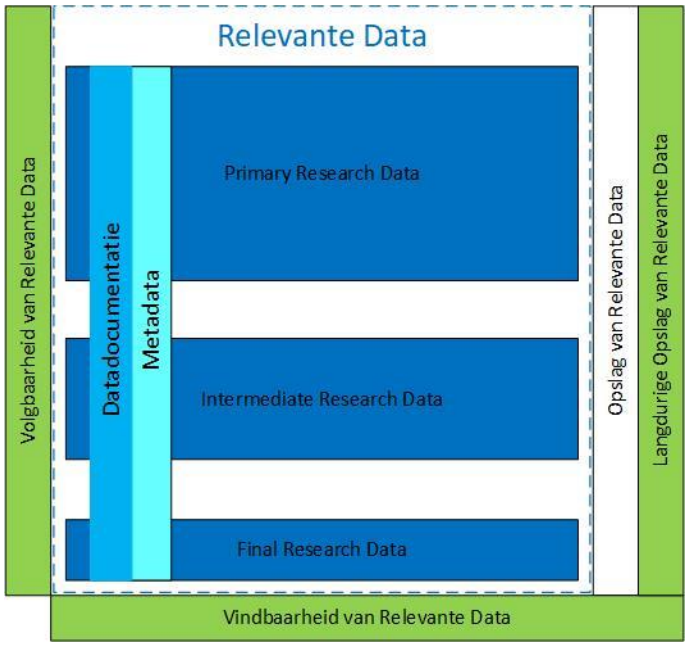
Ontwerp van een RDM werksysteem met de nadruk op vindbaarheid, volgbaarheid en langdurige opslag van onderzoeksdata

Kuipers, L.M.

Award date:
2021

Ontwerp van een RDM werksysteem met de nadruk op vindbaarheid, volgbaarheid en langdurige opslag van onderzoeksdata

Design of an RDM work system with an emphasis on findability, trackability, and long-term storage of research data



Opleiding:	Open Universiteit, faculteit Management, Science & Technology Masteropleiding Business Process Management & IT
Programme:	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	L.M. Kuipers
Identiteitsnummer:	
Datum:	17-06-2021
Afstudeerbegeleider	Wim Penninx
Meelezer	Rik Bos
Versie nummer:	1.0
Status:	definitief

Abstract

Dit onderzoek beoogt antwoord te geven op de vraag: Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op de relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en voor relevante periode op te slaan zijn?

Het onderzoek presenteert daartoe een model in de vorm van een worksystem snapshot dat gebruikt kan worden als globaal ontwerp bij het inrichten van een werkomgeving met mensen, processen en systemen die zo met hun relevante data omgaan dat deze vindbaar, volgbaar en voor relevante periode op te slaan zijn.

Relevante data zijn alle data die nodig zijn om (delen van) het onderzoek te herhalen. Het is over het algemeen een combinatie van onderzoeksdata, metadata en datadocumentatie. Welke data dit precies zijn, wordt bepaald door het nut dat de hoofdonderzoeker voorziet. Deze bepaalt ook de relevante duur van de opslag.

De theorie die opgebouwd is voor dit onderzoek en gebruikt is om het WSS te vullen, is het resultaat van het literatuuronderzoek, semigestructureerde interviews (grounded theory) en een expert review. Het ontwikkelen van het WSS is in drie slagen gedaan, m.b.v. Design Science Research.

Sleutelbegrippen

RDM, werksysteem, vindbaar, volgbaar, onderzoeksdata, metadata, datadocumentatie

Samenvatting

Bij de case organisatie uit dit onderzoek (één van de Nederlandse universiteiten) wordt het volgende geconstateerd:

Alle data van een onderzoek (of onderzoeker) staan verspreid over diverse systemen, vaak ook op diverse locaties, zonder dat zichtbaar gemaakt kan worden welke data bij elkaar horen.

Dit is niet uniek voor de case organisatie; uit literatuuronderzoek blijkt dat het ook voor andere academische omgevingen geldt. Het leidt tot de volgende probleemstelling:

Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op de relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en voor relevante periode op te slaan zijn?

Het gaat in dit onderzoek specifiek om de periode dat het onderzoek gaande is en data worden verzameld en bewerkt. Om de hoofdvraag te beantwoorden wordt een model ('Work System Snapshot' of WSS) gemaakt dat toegespitst is op het vindbaar en volgbaar maken en voor relevante periode op slaan van relevante data.

Het onderzoek wordt in drie fases uitgevoerd en volgt de Design Science Research methode. Eerst wordt theorie opgebouwd met literatuuronderzoek waarna een eerste versie van het WSS gemaakt wordt. Vervolgens wordt de theorie geverifieerd en waar nodig aangevuld door theorie op te bouwen uit veertien semigestructureerde interviews bij de case organisatie met grounded theory. De uiteindelijke versie van de producten en diensten in het WSS volgt uit een expert review volgens de Delphi methode.

De belangrijkste bevindingen van de opgebouwde theorie uit het onderzoek zijn:

- Onderzoekers bij de case organisatie hebben een sterke focus op hun wetenschappelijke probleem; ze willen daarbij zo min mogelijk afgeleid worden door 'randzaken' en een zeer hoge mate van vrijheid hebben in de invulling van hun werkzaamheden, inclusief RDM.
- Relevante data zijn alle onderzoeksdata, metadata en datadocumentatie tezamen, die benodigd zijn voor het herhalen van (delen van) het onderzoek. Wat relevant is wordt bepaald door de hoofdonderzoeker.
- Onderzoeksdata hebben verschillende niveaus (die verlopen van direct na verzamelen tot de data die nodig is voor de tabellen en figuren in de publicatie). Volgbaarheid beschrijft overgangen binnen de niveaus en ertussen en waar dat geregistreerd wordt (in metadata en datadocumentatie). De volgbaarheid van de relevante data wordt voornamelijk gerealiseerd door gebruik te maken van een ELN. Voor codedata specifiek werkt versiebeheer goed, door gebruik te maken van systemen die specifiek bedoeld zijn voor codebeheer (opslag, versiebeheer en delen);
- Vindbaarheid wordt gerealiseerd door vooral folderstructuren en ELN's en in specifieke gevallen ook naamconventies.
- De sterkste oplossing die zowel vindbaarheid als volgbaarheid goed dekt is het gebruik maken van een ELN gecombineerd met een consequent doorgevoerde folderstructuur per onderzoek.
- Voor lopend onderzoek blijkt niet zozeer archivering, maar opslag van belang. Opslag die een relevante periode beschikbaar is (gedurende het onderzoek beschikbaar is en voor een bepaalde periode erna). De locatie van deze opslag is vaak de plek waar men het in eerste

instantie heeft opgeslagen. Data worden pas verwijderd als ze niet meer relevant zijn of als men prioriteit geeft aan de opslag van andere data.

Het hart van het WSS voor de hoofdvraag wordt gevormd door de products and services en de technology:

- Het werksysteem verzekert dat er in het lopende onderzoek voldoende, door de onderzoeker te bepalen, betrouwbare opslagruimte per onderzoek beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.
- Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen onderzoeksdatahouders en een centrale opslag.
- Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie, waarbij versies worden bijgehouden van de verschillende replica's en waarbij een bepaalde versie van een replica teruggesynchroniseerd kan worden.
- Het werksysteem biedt een opslagplaats waar voor een relevante periode de relevante data van een onderzoek voldoende snel opgehaald kunnen worden. Voldoende snel is afhankelijk van hoeveelheden relevante data en de mate waarin ze nog gebruikt worden.
- Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek te ondersteunen waarbij formats in templates beschreven kunnen worden.
- Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op relevante data te kunnen toepassen.
- Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de relevante data zo op te slaan, dat duidelijk is dat ze bij elkaar horen.
- Het werksysteem maakt het mogelijk relevante data en alle afgeleiden hiervan, te vernietigen en logt deze activiteit.
- Het werksysteem levert voorzieningen om metadata te genereren met zo min mogelijk inspanning van de onderzoeker, bij voorkeur geautomatiseerd op basis van een template of algoritmen, maar ook met voldoende ruimte voor eigen invulling.
- Het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via meerdere interfaces (zoals mens-machine, machine-machine).
- Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.
- Het werksysteem bewaakt de integriteit van de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden in logging.
- Het werksysteem kan integreren met bestaande, veelgebruikte, ELN toepassingen.
- Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen.
- Het werksysteem biedt een ELN dat de mogelijkheid heeft om een link (verwijzing, 'shortcut') vast te leggen naar relevante data in het lopende onderzoek.
- Het werksysteem biedt een ELN aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden, zoals het tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.

Summary

At the case organization from this research (one of the Dutch universities) the following is observed: All data of a research (or researcher) are scattered over several systems, often also in several locations, without being able to visualize which data belong to each other.

This is not unique to the case organization, from literature review it appears to be true for other academic environments as well. It leads to the following problem statement:

What design, described in a worksystem snapshot, describes a Research Data Management work system, focused on the relevant data for ongoing research within the case organization, so that these data are findable, trackable, and storeable for relevant periods of time?

This research is specifically about the period when research is ongoing and data is being collected and processed. In order to answer the main question, a model ('Work System Snapshot' or WSS) is made that is focused on making relevant data findable, traceable and storeable for a relevant period.

The research is conducted in three phases and follows the Design Science Research method. First theory is built up by means of literature research after which a first version of the WSS is made. Then the theory is verified and where necessary supplemented by building theory from fourteen semi-structured interviews with the case organization with grounded theory. The final version follows from an expert review using the Delphi method.

The main findings of the built theory from the research are:

- Researchers at the case organization have a strong focus on their scientific problem; they want to be distracted as little as possible by 'peripheral issues' and have a very high degree of freedom in the interpretation of their work, including RDM.
- Relevant data are all research data, metadata and data documentation together, which are needed to repeat (parts of) the research. What is relevant is determined by the principal investigator.
- Research data have different levels (going from immediately after collection to the data needed for the tables and figures in the publication). Traceability describes transitions within the levels and between them and where that is recorded (in metadata and data documentation). Traceability of relevant data is mainly achieved by using an ELN. For code data specifically, versioning works well, using systems specifically for code management (storage, versioning and sharing);
- Findability is achieved primarily through directory structures and ELNs, and in specific cases, name conventions.
- The strongest solution that covers both findability and trackability well is the use of an ELN combined with a consistently implemented folder structure per study.
- For ongoing research, it appears that storage rather than archiving is important. Storage that is available for a relevant period (during the research and for a certain period afterwards). The location of this storage is often where it was initially stored. Data is only deleted when it is no longer relevant or when one prioritizes the storage of other data.

The heart of the WSS for the main question is formed by the products and services and the technology:

- The work system ensures that in the current research sufficient, by the researcher to determine, reliable storage space per research is available to store the relevant data needed to repeat the research, or parts thereof.
- The work system provides functionality to synchronize data (in both directions) between research data holders and a central repository.
- The work system provides automated backup of files and folders by synchronizing (i.e., restores in both directions) with a central storage location, maintaining versions of the various replicas, and allowing a particular version of a replica to be synchronized back.
- The work system provides a repository where, for a relevant period, the relevant data of a study can be retrieved sufficiently quickly. Sufficiently fast depends on quantities of relevant data and the extent to which they are still used.
- The work system offers the possibility to support a free, advised or imposed naming format and folder structure for the relevant data in the current study where formats can be described in templates.
- The work system provides a mechanism to apply automated versioning to relevant data.
- The work system ensures that relevant data from ongoing research are discoverable and interpretable by storing the relevant data in such a way that it is clear that they belong together.
- The work system enables the destruction of relevant data and all derived data and logs this activity.
- The work system provides facilities for generating metadata with as little effort on the part of the researcher as possible, preferably automated on the basis of a template or algorithms, but also with sufficient scope for personal interpretation.
- The work system provides facilities to exchange information and receive assignments through multiple interfaces (such as human-machine, machine-machine).
- The work system offers the possibility to secure the relevant data in the ongoing research in line with the confidentiality of the research.
- The work system monitors the integrity of the relevant data for accuracy and consistency by keeping track of all operations in logging.
- The work system can integrate with existing, commonly used, ELN applications.
- The work system provides a generic ELN that allows logging of activities, ideas, choices, and whatever else is relevant to the research.
- The work system provides an ELN that has the ability to capture a link (reference, 'shortcut') to relevant data in the ongoing research.
- The work system provides an ELN that can be made specific to certain fields, such as drawing molecular structures in chemistry and sample management in biology.

Inhoudsopgave

1. Introductie

In dit hoofdstuk wordt de centrale onderzoeksvraag beschreven. De context wordt gevormd door onderzoek aan een Nederlandse universiteit (hierna onderzoeksinstelling) zoals dat in de periode 2015 – 2021 plaatsvindt. Het hoofdstuk beschrijft de aanleiding, de doelstelling, de relevantie en de centrale onderzoeksvraag van dit afstudeeronderzoek.

1.1. Achtergrond

In twee interviews[1] met twee verschillende wetenschapsondersteuners van verschillende afdelingen bij de case organisatie, werd een belangrijk vraagstuk benoemd:

1. Alle data van een onderzoek (of onderzoeker) staan verspreid over diverse systemen, vaak ook op diverse locaties, zonder dat zichtbaar gemaakt kan worden welke data bij elkaar horen.

In een derde interview[2] met de manager van de archiefrepository van de case organisatie, wordt deze conclusie bevestigd en wordt er één wens aan toegevoegd:

2. Men zou graag alle relevante data van een onderzoek op eenvoudige wijze in de archiefrepository van de case organisatie willen archiveren.

Dit is niet uniek voor de case organisatie, het blijkt ook voor andere academische omgevingen te gelden.

- Het probleem van veel verschillende soorten onderzoeksdata die op veel verschillende locaties opgeslagen worden, wordt in verschillende surveys aan verschillende universiteiten in o.a. Engeland, de Verenigde Staten en Canada beschreven en bevestigd (Akers and Doty 2013; Parsons, Grimshaw, and Williamson 2013; Sewerin et al. 2015).
- Data management, met name het maken van een data management plan, is voor belangrijke subsidieverstrekkers als de NWO een voorwaarde om in aanmerking te komen voor onderzoeksfinanciering. De Europese Unie vaardigt actief richtlijnen uit t.a.v. data management in het Horizon 2020 programma (European Commission 2013). Archivering en traceerbaarheid van data zijn daar aspecten van.

Doelstelling van het onderzoek

De doelstelling van het onderzoek is om tot een globale conceptuele voorstelling te komen van samenwerkende mensen en systemen die alle relevante onderzoeksdata van een onderzoek of onderzoeker vindbaar, volgbaar en archiveerbaar maken en houden.

Leeswijzer

Dit hoofdstuk bevat de achtergrond, de probleemstelling en een globale opzet van de aanpak. Hoofdstuk twee bevat het literatuuronderzoek. In hoofdstuk drie wordt ingegaan op de gebruikte methodes in het empirische deel van het onderzoek en hun verantwoording en in hoofdstuk vier worden de resultaten gepresenteerd. Hoofdstuk vijf bevat de discussie, conclusies en aanbevelingen.

1.2. Gebiedsverkenning

Universiteiten hebben last van versnippering van de opslag van onderzoeksdata. Uit de persoonlijke communicatie met de wetenschapsondersteuners blijkt dat dit probleem gedurende het hele onderzoek speelt. Dit probleem valt binnen het opkomende vakgebied van de Research Data

[1] Persoonlijke communicaties op 22-9-2015 en op 24-9-2015.
[2] Persoonlijke communicatie op 12-4-2017.

Management (RDM) en met name het onderdeel ‘Managing active data’ (Jones, Pryor, and Whyte 2013). Er wordt beoogd de versnippering tegen te gaan door de inzet van een werksysteem, wat bovendien latere eenvoudige archivering van de onderzoeksdata mogelijk maakt. Begrippen die hierbij een rol spelen zijn:

Research data, zijn onderzoeksdata: Burgi, Blumer, & Makhlouf-Shabou (2017) volgen de OECD[3] (Pilat and Fukasaku 2007) in hun definitie van research data wat neerkomt op: feitelijke data die worden gebruikt als primaire bronnen voor wetenschappelijk onderzoek en die algemeen aanvaard zijn in de wetenschappelijke gemeenschap om de onderzoeksresultaten te valideren.

Research Data Management (RDM): Datamanagement is kort samengevat het creëren, opslaan, onderhouden, beschikbaar maken, archiveren en langdurig bewaren van onderzoeksdata. Hierbij wordt als einddoel vaak gerefereerd aan de zogenaamde FAIR principes: Findable, Accessible, Interoperable and Reusable' (Universiteit Leiden n.d.).

Managing active data: Managing active data is een fase in RDM en beslaat de fase van het onderzoek waarin data gecreëerd, bewerkt en geanalyseerd worden. Hierbij moet voldoende opslagruimte in capaciteit, flexibiliteit en functionaliteit aangeboden worden (Burgi et al. 2017; Jones et al. 2013).

Metadata: Metadata zijn data die informatie geven over de onderzoeksdata met als doel de onderzoeksdata voor anderen bruikbaar te maken (reproduceerbaar en interpreteer baar) (National Science Board and National Science Foundation 2005). Bovendien kunnen metadata een indicatie geven van de waarde van de verzamelde gegevens en daarmee mogelijk een juiste herhaling van het experiment bevorderen (validiteit).

Datadocumentatie: “Datadocumentatie tijdens onderzoek betekent het georganiseerd bijhouden van aantekeningen over hoe de data zijn verzameld, wat de resulterende databestanden zijn en hoe ze zijn verwerkt.” (Anon n.d.)

Relevante data: Relevante data zijn alle onderzoeksdata, metadata en datadocumentatie tezamen, die benodigd zijn voor het herhalen van (delen van) het onderzoek. Wat relevant is wordt bepaald door de hoofdonderzoeker.

Data management plan: Een DMP beschrijft de data management levenscyclus voor de data die worden verzameld, verwerkt en/of gecreëerd in een (Horizon 2020) project. Hierbij zijn de data vindbaar (Findable), toegankelijk (Accessible), interopereerbaar (Interoperable, uitwisselbaar tussen verschillende systemen) en herbruikbaar (Reusable) (FAIR) (European Commission 2013).

Databeveiliging: “Bescherming van data tegen ongeautoriseerde (per ongeluk of opzettelijk) wijziging, vernietiging of openbaarmaking.” (Kissel, Locke, and Gallagher, p59, 2011).

Vindbaarheid (voor dit onderzoek): De stakeholders in de managing active data fase van een onderzoek weten waar de data van een onderzoek staat.

Volgbaarheid (voor dit onderzoek): Volgbaarheid omvat de data provenance, het proces van het bijhouden van wijzigingen in de onderzoeksdata, inclusief de middelen waarin die wijzigingen worden bijgehouden.

[3] Organisation for Economic Co-operation and Development

Archiveerbaarheid (voor dit onderzoek): Archiveren kan opgesplitst worden in twee vormen[4]:

1. Het opslaan van relevante data uit het afgeronde onderzoek bij een daarvoor gespecialiseerde repository.
2. Het voor een relevante periode opslaan van relevante data uit het lopende of afgeronde onderzoek op een plek waar deze data later teruggevonden kan worden.

Relevante periode: De duur van de relevante periode om data op te slaan, wordt bepaald door de onderzoeker of onderzoekscoördinator.

Design Science Research (DSR): is een onderzoeksmethode met als doel de doelmatige definitie van een IT 'artifact', dat gemaakt is om belangrijke organisatorische problemen aan te pakken. Het centrale probleem is een zogenaamd 'wicked problem' en de oplossing wordt beschreven door een (design) 'artifact' (Hevner et al. 2004). In dit onderzoek wordt i.p.v. een informatiesysteem een werksysteem (Alter 2008) gebruikt.

Wicked problem: Hevner et al (2004) beschrijven 'wicked problems' samengevat als: Complexe, constant veranderende en onvolledig beschreven (en te beschrijven) problemen waarvan de verschillende aspecten elkaar beïnvloeden. Vaak zijn meerdere oplossingen mogelijk, maar is het onduidelijk welke oplossing het beste is. De aanpak ervan vergt flexibiliteit, creativiteit en teamwork om tot effectieve resultaten te komen.

Artifact of 'design artifact': 'Design is an Artifact', een artefact is een door mensen bedacht object, idee of concept.

Work system (werksysteem): "Een werksysteem is een systeem waarin menselijke deelnemers en/of machines werk (processen en activiteiten) uitvoeren met behulp van informatie, technologie en andere middelen om specifieke producten/diensten te produceren voor specifieke interne en/of externe klanten" (Alter 2008:p6).

Work System Snapshot: "Het WSS is een samenvatting van één pagina van een werksysteem dat de belangrijkste componenten van zes centrale elementen van het werksysteem identificeert"(Alter 2006:p16).

1.3. Probleemstelling

De probleemstelling geeft de centrale vraag weer en de opdrachtformulering deelt deze op in onderzoeksvragen.

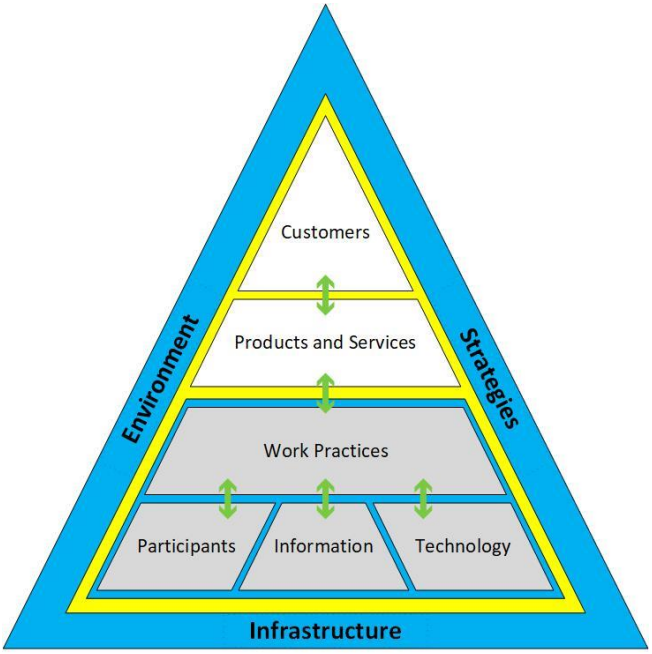
Probleemstelling:
Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op het beheer van actieve onderzoeksdata binnen een Nederlandse universiteit, zodat deze data vindbaar, volgbaar en archiveerbaar zijn?[5]

1.4. Opdrachtformulering

Om tot een oplossing van de probleemstelling te komen, zal eerst onderzocht moeten worden wat het 'managen van active data' precies inhoudt, met name de aspecten die over het vinden en volgen van de onderzoeksdata gaan, en welke soorten onderzoeksdata daar bij horen. Vervolgens wordt er

[4] Tijdens de literatuurstudie wordt met name uitgegaan van het eerste deel van de definitie, tijdens het empirische deel wordt het tweede deel geïntroduceerd en gebruikt.
[5] De probleemstelling zal zich gedurende het onderzoek blijken te ontwikkelen, dit wordt verantwoord en staat beschreven in BIJLAGE ONTWIKKELING VAN DE PROBLEEMSTELLING.

volgens de WSM een snapshot gemaakt van het te ontwikkelen werksysteem. De onderzoeksvragen zijn gebaseerd op 'managing active data' en de toepassing ervan in een werksysteem (FIGUUR 1).



Figuur 1, werksysteem met omgevingsfactoren (Alter 2006).

Dit leidt tot de volgende onderzoeksvragen:

1. Wat is management van active data?
2. Welke data spelen daarbij een rol en wat zijn de eigenschappen van die data?
    a. Categorieën van data;
    b. Hoedanigheden van data;
    c. Dataprincipes.
3. Hoe kan een worksystem snapshot voor het managen van active data op basis van de theorie ingevuld worden?
    a. Wie zijn er bij deze work practices betrokken?
        i. Wie zijn de interne en externe customers?
        ii. Wie zijn de participants aan het werksysteem?
    b. Welke products and services moet het werksysteem leveren?
    c. Welke technology zijn er nodig voor de work practices en data opslag (infrastructure[6])?
    d. Wat zijn de bijbehorende work practices?
    e. Welke information is er binnen het werksysteem benodigd om de work practices te kunnen uitvoeren?

Op basis hiervan kan een eerste artefact, een worksystem snapshot van managing active data worden beschreven.

1.5. Motivatie / relevantie

De praktische relevantie van het onderzoek: Data management, en in het geval van dit onderzoek, vooral het aspect van de versnipperde opslag van onderzoeksdata op verschillende systemen, is een

[6] Infrastructure staat voor bestaande infrastructuur. De TU heeft geen bestaande RDM infrastructuur voor active data management. De technische eisen voor infrastructuur zullen onder technology beschreven worden.

probleem dat speelt bij universiteiten. Alleen als men in staat is alle relevante onderzoeksdata aan het eind van het onderzoek te verzamelen, kan men in staat zijn die data ook te archiveren. Dit laatste is een eis uit het Horizon2020 programma en van daaruit ook een eis voor de wetenschappers van academische instellingen in de EU (European Commission 2013). Het onderzoek beoogt een globaal ontwerp in de vorm van een WSS voor een werksysteem aan te dragen. Het WSS geeft een voorstelling van samenwerkende mensen en systemen die alle relevante onderzoeksdata van een onderzoek of onderzoeker vindbaar, volgbaar en archiveerbaar maken en houden.

De wetenschappelijke relevantie van het onderzoek: Er worden twee methodes gecombineerd om tot het eindresultaat te komen: DSR en de Work System Method (WSM). Dat is één keer eerder gedaan, in een andere vorm en een andere situatie (Truex, Alter, and Long 2010). Er wordt een model voor de beschrijving van de oplossing van het probleem in de vorm van een werksysteem gebruikt (worksystem snapshot). Dat wordt gevuld met de oplossingsmogelijkheden die in de literatuur gevonden worden. Dit model wordt in de empirische fase met semigestructureerde interviews verder verdiept en uiteindelijk onderworpen aan een expert review op basis van de Delphi methode. Dit is een werkwijze die ik niet heb kunnen terugvinden in de literatuur. Ook inhoudelijk levert het model een bijdrage aan de wetenschap. Het vullen van een WSS met bevindingen uit het wetenschappelijke veld van RDM, die specifiek zijn voor vindbaarheid, volgbaarheid en archiveerbaarheid levert een inhoudelijke theoretische bijdrage aan de wetenschap. De design workflow (met de combinatie van de twee methodes) die voor dit onderzoek wordt toegepast is de tweede specifieke bijdrage aan de wetenschap.

1.6. Aanpak in hoofdlijnen

Voor dit onderzoek wordt een combinatie gebruikt van de DSR methodiek van Hevner et al. (2004) en de WSM van Alter (2006). Hevner stelt dat DSR toegepast kan worden om een '**design artifact**' voor een **informatiesysteem** op te leveren dat helpt bij het oplossen van een '**wicked problem**'. Het onderzoek sluit daar op de volgende wijze bij aan:

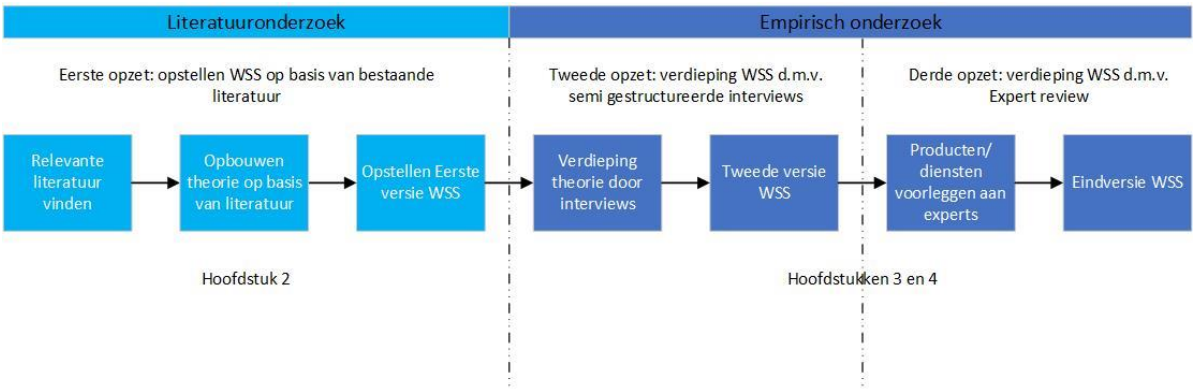
- Een 'design artifact' kan o.a. de vorm aannemen van modellen en van methodes (waar het WSS onder valt).
- RDM kan benaderd worden als een 'wicked problem' (Awre et al. 2015; Cox, Pinfield, and Smith 2016).
  'Managing active data' is een onderdeel van RDM. Gregor en Hevner (2013) stellen dat DSR theorie vooral gericht is op (deel)projecten die onderdeel uitmaken van een groter onderzoeksproject, wat past bij een onderzoek naar het managen van active data in het RDM veld.
- Een informatiesysteem is een bijzondere vorm van een werksysteem (Alter 2008). Werksystemen kunnen via de WSM gemodelleerd worden in een WSS (Alter 2006). De combinatie om een werksysteem te ontwikkelen m.b.v. DSR wordt eerder beschreven door Truex et al. (2010)[7].

Hiermee voldoet dit onderzoek aan de eisen die Hevner et al. (2004) hebben geformuleerd voor DSR.

[7] De opzet van dit onderzoek is anders dan Truex et al. (2010) beschrijven. Op basis van de voorbeelden die zij geven, kan geconcludeerd worden dat ze van bestaande situaties uitgingen, die verbeterd moesten worden. In het geval van de case organisatie is er enigszins sprake van een bestaande situatie. Er zijn veel losse onderdelen van een mogelijke oplossing, maar er is geen samenhangende verzameling werkmethodes en systemen (een werksysteem) die gezamenlijk active data management op zich nemen.

Hevner et al. (2004) stellen voor om eerst het 'problem domain' te beschrijven. Vervolgens gaat er gericht in de literatuur gezocht worden naar mogelijke oplossingen waarna men tot een 'design artifact' komt. In het geval van dit onderzoek worden voor het 'problem domain' de begrippen data (en de kenmerken ervan), vindbaarheid, volgbaarheid en archiveerbaarheid beschreven. De oplossing wordt gezocht in de ontwikkeling van een 'design artifact'.

Modellen kunnen gedefinieerd worden als 'level 2 design artifacts'. (Gregor and Hevner 2013; Hevner et al. 2004). In dit onderzoek worden mogelijke oplossingen voor 'managing active data' in de literatuur gezocht. Deze worden daarna gecontroleerd op hun vindbaarheid, volgbaarheid en archiveerbaarheid. Met de mogelijke oplossingen die voldoen, wordt het WSS (een model) gevuld. Dat kan in een opvolgend empirisch onderzoek verder specifiek gemaakt worden voor een case organisatie. Dit WSS zal ten slotte getoetst worden bij experts. Dit wordt hieronder grafisch voorgesteld in FIGUUR 2.



Figuur 2, Opbouw onderzoek en ontwerp

2. Theoretisch kader

In dit hoofdstuk wordt ingegaan op het zoeken naar literatuur en hoe dat tijdens het lezen en schrijven evolueerde.

2.1. Literatuuronderzoek

Op basis van de eerder genoemde persoonlijke communicatie met de wetenschapsondersteuners en het lezen van artikelen over e-science architectuur (Demchenko et al. 2012) en Electronic Lab Notebooks (ELN) (Oleksik, Milic-Frayling, and Jones 2014; Rubacha, Rattan, and Hosselet 2011) voor het vak 'Document Management' kwam de eerste aanzet voor de probleemstelling tot stand:

Aan welke principes moeten ELN's en onderzoeksdata-opslag voldoen om ze geschikt te maken als middel voor onderzoekers om alle relevante onderzoeksdata van een onderzoek vindbaar, volgbaar en op eenvoudige wijze archiveerbaar te maken binnen een universiteit?

Met name Oleksik et al. (2014) geven aanknopingspunten dat een ELN ingezet kan worden om data in de actieve fase van het onderzoek vindbaar en volgbaar te maken.

Het zoeken naar literatuur gebeurt in eerste instantie in Google Scholar (GS) vanwege de eenvoudige interface en toegankelijkheid en het feit dat het veel resultaten oplevert (niet alleen wetenschappelijk). Daarna worden de zoekopdrachten herhaald in Scopus en in Web of Science (WOS).

Op basis van deze probleemstelling wordt in eerste instantie begonnen met zoeken in de richting van research data en fragmentation en synoniemen daarvan. Dat levert weinig aanknopingspunten op. Uit de persoonlijke communicatie met de wetenschapsondersteuners, bleek dat onderzoeksdata ook op Dropbox stonden. Dat leek typerend voor het probleem, daarom werd het toegevoegd aan "research data" voor de volgende query. Het eerste resultaat was het artikel van Akers en Doty genaamd: Disciplinary differences in faculty research data management practices and perspectives (Akers and Doty 2013). Hierin werd de term "research data management" gebruikt. Na een search hierop bleek dat een veelgebruikt wetenschappelijk begrip te zijn. Vanaf dit moment wordt "research data management" in elke search meegenomen om verankering in het vakgebied te verkrijgen. De eerste set aan queries staat in TABEL 1.

Eerste set aan queries voor het literatuuronderzoek		
NR	Query	Artikelen gevonden? (J/N)
1	research data fragmentation OR shattering OR spread	N
2	"research data" fragmentation OR shattering OR spread	N
3	"research data" dropbox	J
Opmerkingen: Het antwoord in de resultaat kolom is gebaseerd op WOS en Scopus. Dit is veel beperkter dan GS omdat in WOS en Scopus alleen wordt gezocht op peer reviewed artikelen of conference papers.		

Tabel 1, eerste gebruikte queries in het literatuuronderzoek

In het volgende stadium van de literatuurstudie wordt gezocht naar de combinatie van RDM en ELN, het leidt tot de queries uit TABEL 2. Dit levert vooral in GS artikelen op, maar slechts 1 resultaat is Scopus en WOS.

Tweede set aan queries voor het literatuuronderzoek		
NR	Query	Artikelen gevonden? (J/N)
1	“research data management” ELN	J
2	“research data management” “electronic lab notebook”	J
Opmerkingen: Het antwoord in de resultaat kolom is gebaseerd op WOS en Scopus. Dit is veel beperkter dan GS omdat in WOS en Scopus alleen wordt gezocht op peer reviewed artikelen of conference papers.		

Tabel 2, tweede set queries uit literatuuronderzoek

Om het probleemgebied (Hevner et al. 2004) te kwantificeren en backing te krijgen voor de uitspraken in de persoonlijke communicaties met de wetenschapsondersteuners, wordt vervolgens gezocht naar surveys over “research data management” en verschillende vormen van cloud storage. Hiervoor worden de volgende queries uit TABEL 3 gebruikt.

Derde set aan queries voor het literatuuronderzoek		
NR	Query	Artikelen gevonden? (J/N)
1	survey "research data management" dropbox OR "google drive" OR onedrive OR skydrive OR "amazon web"	J
Opmerkingen: De query geeft in GS enkele relevante resultaten, maar het zijn reports en het eerder gevonden artikel van Akers en Doty. Cloud storage wordt vertaald naar dropbox, "google drive", onedrive, skydrive en "amazon web".		

Tabel 3, derde set queries voor het literatuuronderzoek

Er zijn niet veel artikelen die de combinatie ELN en RDM bevatten. Bovendien blijkt uit de surveys dat lang niet alle wetenschappers een ELN gebruiken (Parsons et al. 2013; Sewerin et al. 2015). Het gebrek aan gevonden bruikbare artikelen en mijn dagelijkse werkgebied (Systems en Storage) doen mij in overleg met mijn begeleider besluiten de formulering van de probleemstelling aan te passen:

Aan welke ontwerpprincipes moet een Research Data Management werksysteem, toegespitst op het beheer van actieve onderzoeksdata, binnen een academische omgeving, voldoen zodat deze data vindbaar, volgbaar en archiveerbaar zijn?

Om verdere backing te krijgen voor het gebruik van de DSR methode, wordt ook gezocht op de combinatie van research data management en wicked problem en op de combinatie van research data management en information system. De gebruikte queries staan in TABEL 4.

Vierde set aan queries voor het literatuuronderzoek		
NR	Query	Artikelen gevonden? (J/N)
1	“research data management” “wicked problem”	J
2	“research data management” “information system”	J
Opmerkingen: Het antwoord in de resultaat kolom is gebaseerd op WOS en Scopus. Dit is veel beperkter dan GS omdat in WOS en Scopus alleen wordt gezocht op peer reviewed artikelen of conference papers.		

Tabel 4, vierde set queries voor het literatuuronderzoek

De met de eerste query gevonden artikelen bevestigen dat RDM als een ‘wicked problem’ beschouwd mag worden.

Hoewel de tweede query resultaten oplevert, blijken ze niet van toepassing op het onderwerp van vindbaarheid, volgbaarheid en archiveerbaarheid van actieve onderzoeksdata.

Uit de tot dan toe gevonden artikelen blijkt dat men het heeft over RDM services. Deze services kunnen zowel door (informatie)systemen als door mensen geleverd worden. Deze constatering zal later leiden tot het gebruik van werksystemen i.p.v. information systems, waarbij het artikel van Truex et al. (2010) backing geeft voor het gebruik van de combinatie van DSR en werksystemen.

De laatste query spitst zich toe op de services van RDM (TABEL 5).

Vijfde zoekonderwerp voor het literatuuronderzoek		
NR	Query	Artikelen gevonden? (J/N)
1	"research data management" services.	J
Opmerkingen: Het antwoord in de resultaat kolom is gebaseerd op WOS en Scopus. Dit is veel beperkter dan GS omdat in WOS en Scopus alleen wordt gezocht op peer reviewed artikelen of conference papers. De query geeft resultaten bij Scopus, WOS en GS.		

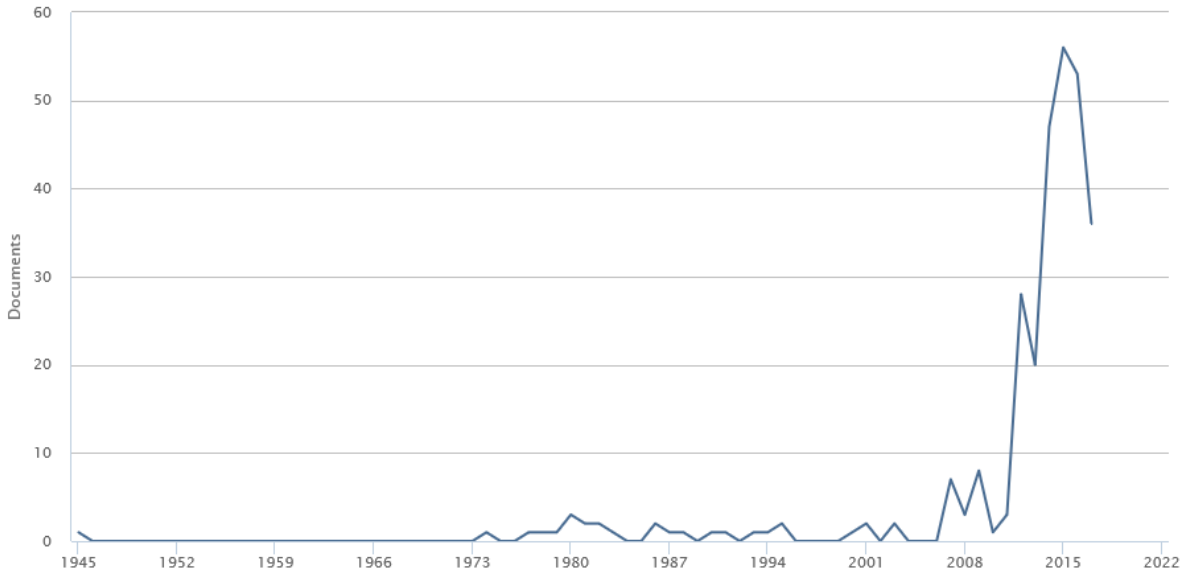
Tabel 5, vijfde zoekonderwerp voor het literatuuronderzoek

Bij het doornemen van de artikelen, blijkt dat het probleemgebied zich bevindt in het onderdeel 'managing active data' van RDM. Als in de eerder gebruikte queries 'research data management' wordt vervangen door "managing active data", levert dat geen nieuwe artikelen meer op.

2.2. Uitvoering

De volgorde van zoeken is in eerste instantie in Google Scholar (GS), Scopus en Web of Science (WOS). Zodra duidelijk wordt dat het vakgebied "research data management" is, kan dit gebruikt worden om de zoekopdracht te verbeteren.

In Scopus kan eenvoudig uitgebeeld worden onder 'analyze search results' hoeveel artikelen per jaar de term "research data management" bevatten (FIGUUR 3). Uit de figuur blijkt dat vanaf 2007 het gebied echte aandacht krijgt in de literatuur. 2007 Blijkt om meer redenen een belangrijk jaar te zijn aangezien Dropbox in dat jaar wordt opgericht (en.wikipedia.org) en uit de persoonlijke communicaties met de wetenschapsondersteuners blijkt dat men hier regelmatig gebruik van maakt om data op te slaan.



Figuur 3,Analyse van zoekresultaten voor "research data management" op www.scopus.com op 23 juli 2017

Bij Scopus kan bij de analyse ook de 'subject area' van de gevonden documenten bekeken worden. De belangrijkste twee voor dit onderzoek zijn 'Computer Science' en 'Social Sciences'. Onder die laatste vallen bij Scopus ook de verschillende 'Library Sciences', die bij WOS apart genoemd staan (in GS, kan geen subject area meegegeven worden aan de search). Voor de zoekopdrachten worden in Scopus daarom de velden computer science, social sciences en library sciences gebruikt. Bij WOS staan deze velden verdeeld over meer categorieën, daarom worden in WOS alle velden aangevinkt

die ‘Computer Science’, ‘Social Science(s)’ of Library bevatten. Omdat de genoemde velden niet allemaal bij GS zijn in te vullen en omdat GS heel veel resultaten geeft die ook minder van toepassing lijken op het vakgebied, wordt, nu het vakgebied duidelijker is geworden, steeds begonnen met zoeken in Scopus, daarna in WOS en als laatste in GS.

Veel van het onderzoek (vooral surveys) gerelateerd aan research data management staat in reports. Omdat zij belangrijke gegevens bieden als het gaat om gebruik van data en middelen om data te creëren, op te slaan, wijzigen etc. worden behalve peer reviewed artikelen en conference proceedings ook reports meegenomen in het literatuuronderzoek. Als taal is gekozen voor Engels en Nederlands. Een overzicht van het voorgaande staat in TABEL 6.

Gebruikte zoekcriteria voor literatuuronderzoek			
Zoekcriterium	Inclusie	Exclusie	Commentaar
Publicatie jaar	2007 en later		Opkomst Dropbox en RDM
Publicatie type	Peer reviewed artikelen, Conference papers Reports		Reports voor de beschrijving van het probleemveld. Vooral surveys over datagebruik
Publicatie vakgebied	“Research data management”		
Journal types	Computer science Library science Social science	Social sciences die niet met library te maken hebben	Library sciences vallen in WOS onder social sciences
Taal	Engels of Nederlands		

Tabel 6, Gebruikte zoekcriteria voor literatuuronderzoek

Beoordeling literatuur en resultaten

Eerst wordt de literatuur op basis van de queries verzameld, met de beperkingen uit TABEL 6. Dit levert uiteindelijk 61 referenties op. In Scopus en WOS kunnen referenties en de ‘geciteerd door’ van een artikel geëxporteerd worden. Dit wordt uitgevoerd en de resultaten worden vervolgens allemaal in een spreadsheet geplaatst. Dit resulteert in bijna 2000 entries. Vervolgens wordt de lijst verkleind door alleen de entries die vaker dan twee keer voorkomen te laten staan. Er blijven 45 entries over. Na het besluit om de probleemstelling aan te passen, vallen de resultaten voor ELN daar af. De resulterende literatuur wordt aan de hand van de samenvatting als volgt beoordeeld:

- De context van het artikel; is het onderzoek gedaan bij Universiteiten of eventueel publieke onderzoeksinstituten?
- Sluit het artikel aan bij de probleemstelling?

Voor aansluiting bij de probleemstelling wordt een waardering tussen matig en zeer goed gegeven. Deze waardering wordt gegeven nadat in ieder geval de samenvatting van het artikel is gelezen. Als dat niet genoeg duidelijkheid geeft voor de waardering, worden eveneens inleiding en conclusies gelezen. Er wordt i.i.g. naar de volgende aanknopingspunten gekeken:

- RDM;
- Managing active data;
- Vindbaar;
- Volgbaar;
- Archiveerbaar.

De laatste drie bullets blijken alleen impliciet in de artikelen te staan, waardoor de waardering in eerste instantie afhangt van de eerste twee bullets en de kwaliteit van de aanknopingspunten die ze bieden (een persoonlijk oordeel op basis van de samenvatting).

De artikelen die redelijk/goed tot zeer goed krijgen, worden in eerste instantie behouden. Er blijven 23 artikelen uit de queries staan en negen resterende referenties. Er zit één artikel in beide lijsten, zodat er 31 overblijven. De lijst die is voortgekomen uit de referenties kan dus wel literatuur van voor 2007 of in een andere taal bevatten (dat laatste komt overigens niet voor). In TABEL 7 staan de aantallen, en of ze zijn opgenomen in de initiële literatuurlijst:

Herkomst gebruikte literatuur			
Soort literatuur	Aantal	Opgenomen in initiële literatuurlijst?	Type publicatie
Primaire bronnen uit originele queries	23	Ja	Journals, reports, conf papers
Artikelen uit de referenties die bruikbaar lijken	9	Ja	Journals
Uit referenties maar ook primaire bron	1	Ja	Journals

Tabel 7, Herkomst gebruikte literatuur

Uiteindelijk blijven er na bestudering dertien artikelen over die in het onderzoek worden opgenomen. De rest van de literatuurlijst komt voort uit andere bronnen. Dat wordt hieronder beschreven.

Extra literatuur voor Design Science Research
De begeleiding heeft in eerste instantie gewezen op het gebruik van DSR van Hevner. Daar wordt gericht op gezocht. Dat leverde twee gebruikte artikelen op.

Via Hevner, maar ook bij gebrek aan goede artikelen over informatiesystemen in combinatie met research data management, wordt de aandacht gevestigd op Work Systems. Die zijn breder georiënteerd dan informatiesystemen en geven via de ‘work system method’ en het ‘work system snapshot’ een goede start om de eerste versie van een artefact op te leveren.

Naar artikelen over Work Systems van Alter hoeft niet echt gezocht te worden want ze staan verzameld op zijn website op researchgate (Alter n.d.); Voor de WSM is zijn boek gebruikt wat is gebaseerd op zijn eerdere research (Alter 2006). Alter is de bedenker van het ‘work system’ en de ‘work system method’ en schrijft er veruit het meest over. Uiteindelijk worden er, met het boek meegerekend, drie referenties van Alter gebruikt.

Door te zoeken op design science research en work system method wordt het artikel van Truex et al. (2010) gevonden.

Extra literatuur die werd gevonden tijdens het schrijven
Gedurende het onderzoek zijn er nog artikelen bijgekomen na het lezen van eerder gevonden artikelen. Dit gebeurde als verduidelijking of verdieping noodzakelijk was op een bepaald onderwerp.

- Bij het categoriseren van data maakten Burgi et al. (2017) veel gebruik van Borgman. Het boek van Borgman is gebruikt om bij die categorisering duidelijke informatie te geven (Borgman 2017). Hetzelfde geldt voor de National Science Board, die bovendien verduidelijking in de definitie van metadata toevoegt (National Science Board and National Science Foundation 2005).

- Bij het zoeken naar een plaatje voor de scientific data lifecycle werd de RDM website van de universiteit van Leiden gevonden alsook het artikel van Verhaar et al. Beiden bleken goed aan te sluiten bij het onderzoek (Universiteit Leiden n.d.; Verhaar et al. 2017). Ook Rebouillat verwijst naar de website van Leiden (Rebouillat 2017);
- De research data management pyramid van Martin Lewis (Lewis 2010) wordt in het artikel van Cox en Pinfield aangehaald (Cox and Pinfield 2014). In eerste instantie wordt naar meer artikelen van Lewis gezocht omdat de pyramid een mooi RDM overzicht geeft. Hiervoor wordt de zoekterm: "research data management" author:lewis in GS gebruikt. Het levert een resultaat op van een andere Lewis (John). Een report met heel veel informatie over active data management (en RDM in het algemeen). Het geeft zeer specifieke beschrijvingen van wat er nodig is in de actieve fase van het onderzoek (Lewis and A. 2014). De research data management pyramid van Martin Lewis haalt het artikel uiteindelijk niet omdat het niet specifiek genoeg in de active data management hoek zit;
- In geen van de artikelen stond een goed overzicht van de stakeholders binnen RDM. Er werd gezocht naar “research data management” stakeholders. Na enig zoeken werd het artikel van Flores et al. gevonden, waar een groot deel van paragraaf 2.7 over customers en participants op gebaseerd is (Flores et al. 2015).
- In hun artikel verwijzen Jones et al. (2013) naar het artikel van Monash University en RDM inspanningen aan de universiteit van Edinburgh (Jones 2013). Na zoeken op RDM en Edinburgh wordt het artikel van Rice et al. (2013) gevonden.
- Het artikel van Starr et al. (2015) voegt diepte toe aan de definitie van metadata en de minimale vereisten ervan.
- De artikelen van Demchenko et al. (2012), Oleksik et al. (2014) en Rubacha et al. (2011) hebben de eerste richting gegeven voor de uiteindelijke probleemstelling tot stand kwam.

De uiteindelijke aantallen staan in TABEL 8.

Overzicht aantallen gebruikte literatuurbronnen		
Bron	Gebruikt	Waarvan summier gebruikt[1]
Journal	15	3
Conference	6	2
Book	3	1
Report	5	0
Web	10	0
[1] summier houdt in dit geval in dat de bronnen slechts zijdelings zijn gebruikt.		

Tabel 8, overzicht aantallen gebruikte literatuurbronnen

In BIJLAGE LITERATUURLIJST staat de lijst gevonden bronnen en de reden waarom ze uiteindelijk wel of niet gebruikt zijn.

2.3. Resultaten en conclusies

Hieronder wordt theorie opgebouwd en worden de onderzoeksvragen voor de literatuurstudie beantwoord. De eerste drie paragrafen beantwoorden de eerste twee onderzoeksvragen. De paragrafen 2.7 t/m 2.7.5 zijn specifiek voor het vullen van het WSS. Paragraaf 2.7.6 ten slotte, toont het gevulde WSS en beantwoord daarmee de derde vraag. De vragen staan hier nog kort benoemd:

1. Wat is management van active data?
2. Welke data spelen daarbij een rol en wat zijn de eigenschappen van die data?
3. Hoe kan een worksystem snapshot voor het managen van active data op basis van de theorie ingevuld worden?

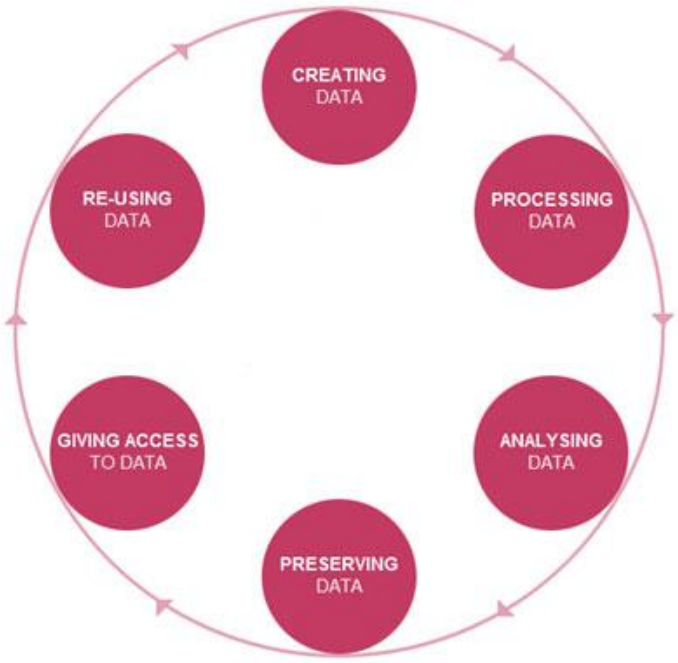
2.4. Managing of active data

Een onderdeel van RDM is 'managing of active data' (Burgi et al. 2017; Jones et al. 2013). Burgi et al. (2017) definiëren actieve onderzoeksdata als de data tijdens de fases van 'collecting', 'processing' en 'analysing'. Volgens Jones et al. (2013) zijn de twee belangrijkste concerns t.a.v. het managen van (onderzoeks)data tijdens de actieve fase van onderzoek:

1. Voldoende opslagruimte aanbieden.
2. Mogelijkheden bieden aan onderzoekers om hun data op functionele en flexibele wijze op te slaan, te benaderen en te delen.

Burgi et al. (2017) geven op basis van het onderzoek van Klindt & Amrhein (2015) een derde concern aan, namelijk het kunnen interpreteren van de data op elk moment in het onderzoek. Data transformeren tijdens een onderzoek door de verschillende bewerkingen die er op verricht worden. Na elke bewerking is het noodzakelijk dat de data interpreteerbaar blijven (volgbaar).

Er zijn diverse modellen die de verschillende fases die de data doorlopen tijdens een onderzoek beschrijven. Het model van de UK Data service wordt in Nederland gebruikt door de Nederlandse Federatie van Universitair Medische Centra (NFU) en de universiteit van Leiden (NFU n.d.; UKDataService 2016; Verhaar et al. 2017) en staat in FIGUUR 4.



figuur 4, scientific data lifecycle model volgens de UK Data Service (UKDataService 2016)

Bij de universiteit van Leiden splitst men het onderzoek, t.b.v. datamanagement, op in drie fases waarbij met name de 'tijdens' fase voor dit onderzoek relevant is:

> **"Tijdens**: opslag van data (locatie, back-up), naamgeving, (mappen)structuur en versiebeheer, metadata en documentatie (de beschrijving van/uitleg over de data) en de toegang tot data (wie er wel of niet bij mag en met welke autorisaties tijdens het onderzoek)" (Universiteit Leiden n.d.).

Lewis heeft het over ‘once underway’ als hij de fase van een research project beschrijft waarin men met active data werkt. Hij benoemt het als de fases waarin ruwe data worden gecreëerd/verzameld, diverse keren worden bewerkt en uiteindelijk tot resultaten leiden (de resultaten kunnen ook weer een dataset zijn) (Lewis and A. 2014). Op basis hiervan kan gesteld worden dat ‘management of active data’ overeenkomt met de activiteiten in de ‘tijdens’ fase zoals de Universiteit van Leiden die benoemt. Het gaat om de onderdelen creating, processing en analysing data uit de scientific lifecycle (FIGUUR 4).

2.5. Theorieopbouw onderzoeksdata

In eerste instantie wordt ingegaan op de categorisering van de onderzoeksdata. Daarna wordt op basis van enkele surveys vastgesteld, in welke hoedanigheden de data in de praktijk voorkomen op verschillende universiteiten. Er wordt afgesloten met data principes. Tezamen vormen ze de theorie rondom onderzoeksdata.

2.5.1. Categorieën van onderzoeksdata

Burgi et al. (2017) volgen de OECD (Pilat and Fukasaku 2007) in hun definitie van research data: ‘feitelijke data (zoals numerieke scores, tekstuele records, afbeeldingen en geluiden) die worden gebruikt als primaire bronnen voor wetenschappelijk onderzoek en die algemeen aanvaard zijn in de wetenschappelijke gemeenschap om de onderzoeksresultaten te valideren’.

Onderzoeksdata hebben twee belangrijke eigenschappen:

1. Ze dienen als primaire bron voor wetenschappelijk onderzoek;
2. In de wetenschappelijke gemeenschap worden ze geaccepteerd als noodzakelijk om wetenschappelijke bevindingen te valideren (Burgi et al. 2017; Pilat and Fukasaku 2007).

De tweede stelling verwoordt dat de wetenschappelijke gemeenschap onderzoeksdata noodzakelijk vinden om onderzoek te valideren. Alleen de wetenschapper die het onderzoek uitvoert, kan bepalen welke data relevant zijn om het onderzoek te valideren. Een derde stelling wordt daarmee:

3. De onderzoeker bepaalt wat de relevante data van een onderzoek zijn.

Burgi et al. (2017) en Borgman (2017) volgen de US National Science Board (2005) door de data onder te verdelen in: ’observational, computational or experimental’ data. Borgman voegt daar de categorie ‘records’ aan toe:

- Observational data: data die voortkomen uit het herkennen en vastleggen van feiten over en/of optreden van bepaalde geobserveerde gebeurtenissen/verschijnselen.
- Computational data: data die voortkomen uit het uitvoeren van computermodelle en/of simulaties.
- Experimental data: data die voortkomen uit het onderzoeken van gebeurtenissen in gecontroleerde omstandigheden om een hypothese te testen, of om nieuwe wetten te ontdekken of te testen.
- Records: data die niet voortkomen uit voornoemde categorieën, het gaat om allerlei soorten ‘vastleggingen’ die voor wetenschappelijk onderzoek kunnen dienen. Beeld- en geluidsfragmenten, boeken, maar ook ELN’s (Borgman 2017; Burgi et al. 2017).

Burgi et al. (2017) maken op basis van het onderzoek van Klindt and Amrhein (2015) ook nog onderscheid tussen de opslag van passieve en actieve data.

- Passieve data die niet meer veranderen kunnen eenvoudig op bitniveau worden opgeslagen en
- Actieve data moeten na elke transformatie interpreteerbaar blijven en per versie worden opgeslagen.

Voor dit onderzoek gaat het om actieve onderzoeksdata uit alle categorieën, inclusief de data die nodig zijn om ze interpreteerbaar te houden.

2.5.2. Hoedanigheden van onderzoeksdata

Er zijn drie verschillende surveys bij universiteiten bestudeerd om datakarakteristieken als locatie, hoeveelheid en soort te bepalen. Dit staat samengevat in TABEL 9.

Datakarakteristieken (locatie, hoeveelheid, soort)				
Data karakteristiek	voorbeelden	kwantiteit	bron	opmerkingen
Datasoorten	ruwe data, applicatie data, documenten	>17 soorten	1, 2 en 3	meerderheid documenten en spreadsheets
Bestandsformaten	Docx, pdf, xlsx	Veel (een aantal per datasoort)	1, 2 en 3	docx, pdf
Opslaglocaties*	laptops, meetpc's, usb drives	>19 locaties	1, 2 en 3	58% (2) tot 67% (3) decentrale systemen
Opslaghoeveelheid	Geen	enkele MB's tot honderden TB's	1, 2 en 3	
Backuplocaties	Centrale servers, usb drives, dropbox	>19 locaties	1 en 3	53% (3) staat decentraal
Backupfrequentie	Dagelijks, wekelijks	35% dag, 16% week, 38% ad hoc	3	9% weet niet wanneer en 25% doet het nooit

Opmerkingen:
Bronnen:
1. (Alexogiannopoulos, McKenney, and Pickton 2010), report, survey.
2. (Buys and Shaw 2015), report ,survey.
3. (Parsons et al. 2013), report, survey.

*Leiden geeft een overzicht van alle mogelijke opslaglocaties die intern en extern gebruikt kunnen worden. In het artikel van Verhaar uit 2017 zijn dat er 44, waarvan er 17 geschikt werden geacht voor de 'tijdens' fase (Verhaar et al. 2017). Het grote aantal verschillende opslaglocaties is een risico voor de vindbaarheid van data.

Tabel 9, datakarakteristieken (locatie, hoeveelheid, soort)

2.6. Data principes

Het Horizon 2020 programma van de EU (European Commission 2013) geeft kaders voor het omgaan met wetenschappelijke data. In het Horizon 2020 programma volgt de EU de FAIR data principes (European Commission 2013; Wilkinson et al. 2016). Burgi et al. (2017) en Awre et al. (2015) refereren naar de OECD data principes (Pilat and Fukasaku 2007). Veel van de principes van de OECD zijn van toepassing op de data die worden opgeslagen (gearchiveerd) na publicatie. Een deel kan ook gebruikt worden voor managing active data (de 'tijdens' fase).

De FAIR principes zijn vooral gericht op het hergebruiken van data en vooral bedoeld om de collectiefase te vergemakkelijken (Wilkinson et al. 2016). Er wordt nadruk gelegd op 'machine actionability' waarbij software data vindt en interpreteert. De FAIR principes leggen daarom veel nadruk op metadata. Het zijn vier principes, die ieder weer onderverdeeld zijn (Findable, Accessible, Interoperable, Reusable).

De OECD principes zijn gericht op de maatregelen die dienen te worden genomen om de toegang tot de data te kunnen regelen. De FAIR en OECD principes staan in TABEL 10 benoemd met de controle

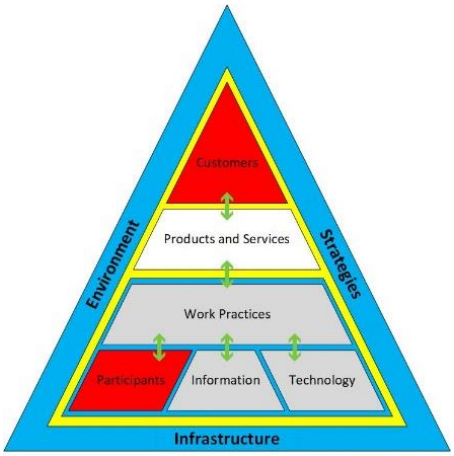
op vindbaarheid, volgbaarheid en archiveerbaarheid. De accessible principes blijken geen relatie met de probleemstelling te vertonen.

FAIR en OECD principes toegepast op de actieve fase met de aspecten vind-, volg- en archiveerbaar				
Van toepassing zijnde aspecten probleemstelling	*VDB[1]*	*VGB[2]*	*ARB[3]*	Commentaar
Vindbaar (OECD: findable)				
Data worden met rijke metadata beschreven	X	X		Volgt definitie metadata.
(meta)data zijn vindbaar in een doorzoekbare bron	X			Duidelijk zonder verdere uitleg
Interopereerbaar (interoperable) (OECD: interoperability)				
(meta)data gebruiken een formele, toegankelijke, gedeelde en wijds beschikbare taal voor kennisrepresentatie	X	X		Volgt definitie metadata.
(meta)data gebruiken vocabulaires die de FAIR principes volgen	X	X		Volgt definitie metadata.
Herbruikbaar (reusable) (OECD: Transparancy, quality, professionalism)				
(meta)data zijn rijk beschreven met voldoende relevante attributen	X	X		Volgt definitie metadata.
(meta)data worden met een duidelijke en toegankelijke dataovereenkomst beschikbaar gesteld			X	Dit geldt voor te archiveren data.
OECD principes die niet (volledig) overlappen met de FAIR principes				
Openness; toegang voor de internationale research gemeenschap	X			De gemeenschap moet daarvoor kunnen vinden in (voor dit onderzoek) de actieve fase.
Security; het waarborgen van de veiligheid en integriteit van de data bij het verlenen van toegang aan derden.		X	X	Kunnen zien dat de oorspronkelijke dataset onveranderd is, geldt voor en na de actieve fase.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid [3]ARB: Archiveerbaarheid				

Tabel 10, FAIR en OECD principes toegepast op de 'tijdens' fase met de aspecten vind-, volg- en archiveerbaar (Pilat and Fukasaku 2007; Wilkinson et al. 2016)

2.7. Het RDM werksysteem

In de paragrafen hierna wordt het WSS gevuld op basis van de hiervoor benoemde theorie en de uitwerkingen in BIJLAGE SAMENVOEGING VAN DE RESULTATEN VOOR PRODUCTS AND SERVICES EN TECHNOLOGY T.A.V. VINDBAARHEID, VOLGBAARHEID EN ARCHIVEERBAARHEID.



Figuur 5, customers en participants voor WSS

2.7.1. stakeholders van het werksysteem

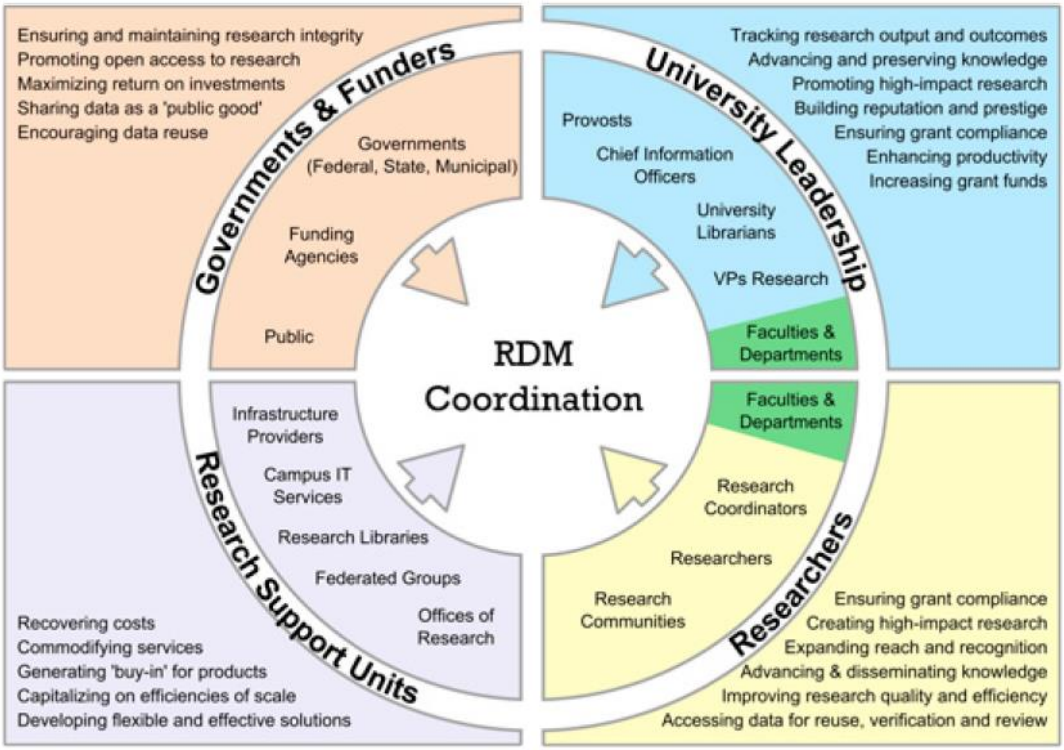
In deze paragraaf worden de customers en de participants van het WSS bepaald. Een work system bestaat om products and services te produceren voor de customers (Alter 2006). Voor het work system in dit onderzoek zijn de customers de onderzoekers van een universiteit die opslag voor hun actieve onderzoeksdata nodig hebben of in gebruik hebben. Participants zijn participants aan het werksysteem. In het geval van selfservice wordt de customer ook een participant (Alter 2006). Dit kan de IT'er zijn die opslagruimte voor het onderzoek reserveert, maar het kan ook de hoogleraar zijn die toestemming geeft voor de uitgaven die gemoeid zijn met de opslag.

Om de customers en de participants te bepalen worden de stakeholders van RDM nader onderzocht. Cox et al. (2016) beschrijven in hun artikel de uitkomsten van 26 semigestructureerde interviews

met Britse bibliothecarissen van hogere onderwijsinstellingen. Ze zijn o.a. op zoek naar de stakeholders binnen RDM en hun perspectief t.a.v. RDM. Hieruit blijkt dat er niet één bepaalde stakeholdergroep specifiek voor de dienst te noemen is. Er zijn verschillende stakeholders met verschillende vaardigheden die allemaal een deel van een nog incompleet antwoord hebben. De ene stakeholdergroep hecht meer belang aan een bepaald aspect van RDM dan de andere.

Flores et al. (2015) geven een overzicht van RDM stakeholders waarbij ze ze in vier categorieën verdelen zoals is te zien in FIGUUR 6. De genoemde stakeholders in de artikelen van Cox et al. en Flores et al. zijn vergelijkbaar, hoewel ze niet dezelfde namen gebruiken. Wilkinson et al. (2016) spreken van een extra stakeholder: de computational stakeholder.

Niet alle genoemde stakeholders hebben te maken met de 'tijdens' fase van het onderzoek. Flores et al. (2015) laten in tabel 1 in hun artikel zien wie de belangrijkste stakeholders zijn bij een specifieke datamanagement service. Dit is hieronder in TABEL 11 overgenomen met daarin alleen de services die specifiek horen bij managing active data, waarbij de activiteiten van de 'tijdens' fase, zoals benoemd in 2.4, zijn vergeleken met die in de tabel van Flores et al.



Figuur 6, Stakeholders onderverdeeld in vier groepen, met de individuele stakeholders in de binnenste ring en de belangen van de groep in de buitenste ring (Flores et al., 2015).

Stakeholders en de daarbij behorende onderdelen van managing active data		
Service	**Key Stakeholders**	**Library's Coordination Role**
Access control	Researchers, research support units	Advise on data embargoing and access control issues
Data documentation	Researchers, research support units	Help researchers determine how best to document their data at the beginning of a project, following disciplinary standards
Intellectual property and copyright	Researchers, research support units	Provide guidance on intellectual property and copyright matters surrounding research data
Preservation	Researchers, research support units	Advise on appropriate data formats for preservation, preparing data sets for long-term preservation
Privacy and confidentiality	Researchers, research support units	Advise researchers and research office staff on privacy and confidentiality issues in data management
Repository selection[8]	Researchers, research support units	Help individuals select trusted digital repositories for preserving data sets, whether those are disciplinary repositories or institutionally managed repositories *and to select trusted digital repositories to store active data during a research project.*

Tabel 11, Stakeholders en de daarbij behorende onderdelen van managing active data (Flores et al. 2015)

De belangrijkste stakeholders zijn de researchers en de research support units.
De research support units (RSU) bestaan uit

- Research libraries;
- Campus IT services;
- Infrastructure providers (mogelijk samenwerkende aanbieders van technische infrastructuur voor RDM diensten);
- Federated groups (samenwerkingsverbanden van (onderdelen van) verschillende instituten om een (deel van een) RDM dienst aan te bieden);
- Offices of research.

RSU kan gezien worden als **participant** in het werksysteem. Zij zijn geen afnemer van de RDM dienst.
De researchers stakeholders (RS) categorie bestaat uit:

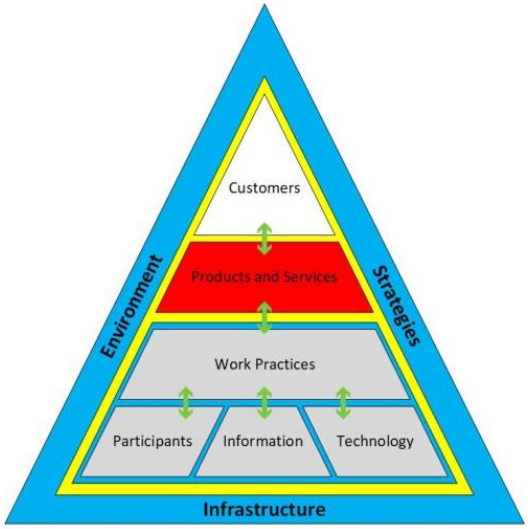
- Research communities (onderzoeksgemeenschappen, vaak ook over instituten heen);
- Researchers (onderzoekers en master studenten[9]);
- Research coordinators (onderzoekscoördinatoren);
- Faculties and departments (faculteiten en afdelingen van faculteiten).

[8] In het artikel van (Flores et al. 2015) wordt het uitzoeken van een repository voor archivering bedoeld. Dat is een taak die idealiter in de actieve fase van het onderzoek al wordt verricht. Het overzicht aan repositories van Leiden, laat bovendien zien dat men ook in de actieve fase van verschillende repositories gebruik kan maken vandaar de cursief getypte toevoeging in de tabel (Universiteit Leiden n.d.).

[9] Studenten in de afstudeerfase worden niet als stakeholder genoemd voor het actieve deel van het onderzoek. De praktijk is daarentegen dat master studenten regelmatig uitvoerenden zijn in een onderzoek. Hun hoedanigheid is dan wellicht meer onderzoeker dan student, zij zouden dan in de researchers groep vallen.

De researchers zijn zowel interne als externe klant van het systeem (**internal** en **external customers** in het werksysteem). Intern binnen het instituut en extern van daarbuiten (bijvoorbeeld een wetenschapper van een andere universiteit waarmee wordt samengewerkt). Bovendien zijn de researchers ook **participants**. Zij leveren o.a. informatie aan om het werksysteem te laten functioneren.

2.7.2. Products and services van het werksysteem



Figuur 7, products and services voor WSS

Om het products and services deel van het WSS te vullen, is het van belang de diensten die het werksysteem biedt te verzamelen (de output). Hieronder staat een opsomming van active data management diensten uit de verschillende artikelen. Om te bepalen of de stelling die in de literatuur gevonden werd een dienst en/of product betreft, wordt de volgende vraag gehanteerd: Gaat het om iets waar een gebruiker om zou kunnen vragen, iets wat hij/zij kan gebruiken? De resulterende opsomming wordt vervolgens beoordeeld op de aspecten vindbaarheid, volgbaarheid en archiveerbaarheid.

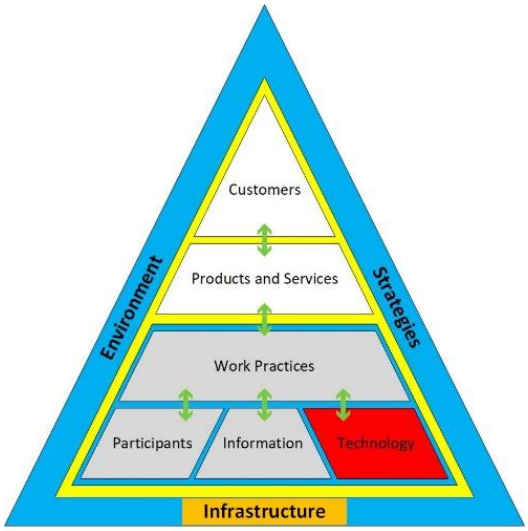
In TABEL 12 staat uitgewerkt hoe de products and services zich verhouden tot de aspecten vindbaarheid, volgbaarheid en archiveerbaarheid.

Products and services afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid				
Product en/of service	VDB[1]	VGB[2]	ARB[3]	Commentaar
Het werksysteem verzekert dat er gedurende het onderzoek voldoende betrouwbare opslagruimte beschikbaar is om de data op te slaan (Cox and Pinfield 2014; Demchenko et al. 2012; Jones et al. 2013; Lewis and A. 2014; Rice et al. 2013; Sewerin et al. 2015)(Repository selection uit TABEL 11);	X			Bekende locatie maakt vinden makkelijker.
Het werksysteem levert sync and share functionaliteit ('academic dropbox') (Alexogiannopoulos et al. 2010; Buys and Shaw 2015; Jones et al. 2013; Lewis and A. 2014; Parsons et al. 2013; Rice et al. 2013);	X			Het syncen naar één centrale locatie verhoogt de vindbaarheid van de mogelijk verspreide data.
Het werksysteem biedt de mogelijkheid dat data van electronic lab notebooks (ELN) verplaatst worden naar centrale active data opslag (Lewis and A. 2014).	X			Als de data op 1 locatie staan, wordt vindbaarheid vereenvoudigd.
Het werksysteem maakt het mogelijk om de data van elk device (pc's, laptops, tablets, telefoons, laboratoriuminstrumenten enz.) te kopiëren (Lewis and A. 2014; Rice et al. 2013);	X			Als de data op 1 locatie staan, wordt vindbaarheid vereenvoudigd.
Het werksysteem biedt de mogelijkheid om data naar meerdere locaties/systemen te distribueren/repliceren. Rice et al. noemen specifiek HPC omgevingen (Demchenko et al. 2012; Rice et al. 2013);	X			Op voorwaarde van geschikte logging blijven de data vindbaar.
Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden (Verhaar et al. 2017).	X			De structuur vereenvoudigt vinden.

Product en/of service	VDB[1]	VGB[2]	ARB[3]	Commentaar
Het werksysteem verzekert dat onderzoeksdata vindbaar en begrijpelijk zijn door de data in combinatie met de bijbehorende metadata en andere documentatie op te slaan (Cox and Pinfield 2014; Lewis and A. 2014; Sewerin et al. 2015; Verhaar et al. 2017)(data documentation in TABEL 11);	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
Het werksysteem verzekert dat backupmogelijkheden gedurende het onderzoek beschikbaar zijn (Lewis and A. 2014; Sewerin et al. 2015; Verhaar et al. 2017);	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.
Het werksysteem zorgt dat het aanmaken van metadata bij voorkeur geautomatiseerd plaatsvindt (Jones et al. 2013; Lewis and A. 2014);	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en voor machines (bijvoorbeeld een API). Dit is noodzakelijk om met het werksysteem te kunnen werken.	X	X		Om data te kunnen vinden en volgen is een interface om mee te zoeken noodzakelijk.
Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen (Alexogiannopoulos et al. 2010; Jones et al. 2013; Lewis and A. 2014; Verhaar et al. 2017);		X		Helpt datatransformaties inzichtelijk te maken.
Het werksysteem levert de mogelijkheid om data te vernietigen (Lewis and A. 2014);		X		Het is de laatste bewerking op de dataset. Voorwaarde is voldoende logging.
Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan (Sewerin et al. 2015; Verhaar et al. 2017);		X		Op voorwaarde van geschikte logging volgbaar.
Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering) (Lewis and A. 2014);		X	X	Volgbare transformatie en klaarmaken voor archief.
Het werksysteem kan unieke identifiers aanmaken voor een dataset (Persistent Identifiers, PID) (Lewis and A. 2014).			X	Een PID* is nodig voor archiveerbaarheid.
Het werksysteem verzekert dat de data selectief beschikbaar gemaakt kunnen worden aan anderen (Lewis and A. 2014; Sewerin et al. 2015; Verhaar et al. 2017) (acces control in TABEL 11);				Dit zegt niets over 1 van de 3 aspecten.
Het werksysteem moet dermate flexibel zijn, dat het een grote variëteit aan gebruik kan ondersteunen (Rice et al. 2013);				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem draagt er zorg voor dat actieve onderzoeksdata snel toegankelijk zijn. Het kan sterke rekenkracht (computational resources) nodig hebben (Demchenko et al. 2012; Lewis and A. 2014).				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem ondersteunt langdurig lopende experimenten (Demchenko et al. 2012)				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem moet privacyeisen, copyrighteisen en eisen uit wet- en regelgeving kunnen waarborgen (Cox and Pinfield 2014; Sewerin et al. 2015; Verhaar et al. 2017) (Intellectual property and copyright en Privacy and confidentiality in TABEL 11);				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem biedt opslag voor de RDM infrastructuur die gelaagd moet zijn naar prijs en daarbij behorende veiligheid van de data (goedkoop en relatief onveilig voor reproduceerbare data en duur en veilig voor niet reproduceerbare data, zie 2.5);				Geen directe aanknopingspunten met één van de drie aspecten.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid [3]ARB: Archiveerbaarheid *Een PID is een persistent identifier, een unieke code waarmee de dataset geïdentificeerd kan worden.				

Tabel 12, Products and Services afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid

2.7.3. Technology en infrastructure van het werksysteem



Figuur 8, technology en infrastructure voor WSS

In deze paragraaf wordt het veld technology van het WSS gevuld. Dit zijn alle technische middelen die gebruikt worden om de work practices in het WSS uit te voeren. Er worden echter vooral eisen aan de infrastructuur[10] gevonden in de literatuur. Met de infrastructure in het WSS wordt bestaande infrastructuur bedoeld die men voor de uit te voeren work practices kan inzetten (Alter 2006). Voor dit WSS wordt technology breder geïnterpreteerd naar technische middelen en eisen aan de op te bouwen infrastructure voor het werksysteem. In de empirische fase zal geëvalueerd moeten worden of deze vrijheid genomen mag worden. Ook deze lijst van uitspraken uit de verschillende artikelen zal weer beoordeeld worden op de aspecten vindbaarheid, volgbaarheid en archiveerbaarheid. De resultaten staan hieronder in TABEL 13.

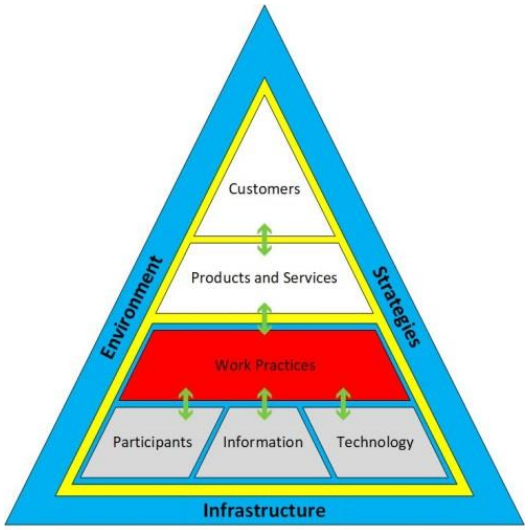
RDM infrastructure voorwaarden afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid				
RDM infrastructure voorwaarde	VDB[1]	VGB[2]	ARB[3]	Kort Commentaar
Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface (Lewis and A. 2014).	X			Op voorwaarde dat de register/index voor doorzoekbaarheid zorgt, een vindbaarheid stelling.
Het werksysteem biedt opslag voor de RDM infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd) (Burgi et al. 2017).	X			Fysieke vindbaarheid, weten waar de data daadwerkelijk staan.
Het werksysteem biedt opslag voor de RDM infrastructuur die schaalbaar is (Burgi et al. 2017).				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem biedt opslag voor de RDM infrastructuur die zelfcorrigerend is (fouten op bitniveau kunnen herstellen) (Burgi et al. 2017);				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem biedt opslag voor de RDM infrastructuur die robuust is (goed tegen (externe) verstoringen kunnen) (Burgi et al. 2017);				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem biedt opslag voor de RDM infrastructuur die kleine tot grote hoeveelheden opslagobjecten per onderzoek kan verwerken (paragraaf 2.5);				Geen directe aanknopingspunten met één van de drie aspecten.
Het werksysteem biedt opslag voor de RDM infrastructuur die veel verschillende bestandsformaten aan kan (paragraaf 2.5);				Geen directe aanknopingspunten met één van de drie aspecten.

[10] In de literatuur spreekt men over RDM (technische) infrastructuur, dat is wat in de opsomming wordt bedoeld, een RDM infrastructuur voor dit WSS.

RDM infrastructure voorwaarde	VDB[1]	VGB[2]	ARB[3]	Kort Commentaar
Het werksysteem biedt verschillende lagen van beveiliging voor wat betreft de toegang tot de data namelijk: via een beveiligde verbinding buiten de campus (sshfs en/of webdav) en via standaard sharing protocollen binnen de campus (CIFS en NFS) (Rice et al. 2013);				Geen directe aanknopingspunten met één van de drie aspecten.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid [3]ARB: Archiveerbaarheid				

Tabel 13, RDM infrastructure voorwaarden afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid

2.7.4. Work practices van het werksysteem



Figuur 10, work practices voor het WSS

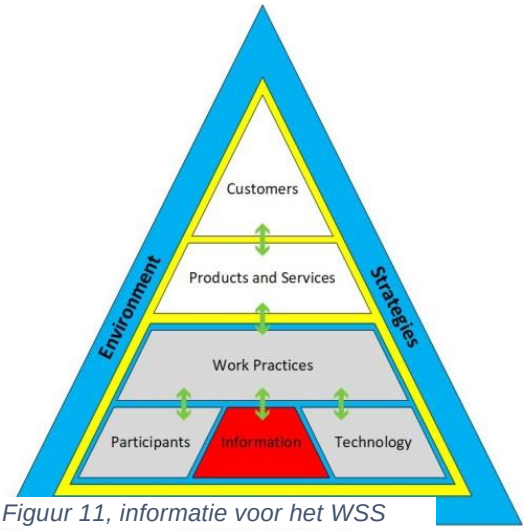
Op basis van het voorgaande kunnen de work practices die bij managing active data horen, vastgesteld worden. De opsomming is bedoeld om het veld work practices in het worksystem snapshot in te vullen. Bij deze opsomming dient men in acht te nemen dat het voorgaande niet specifiek genoeg is om een precieze set aan work practices te beschrijven. De beschreven work practices, sluiten aan op de literatuur en op de te leveren products and services. Ze zijn met name bedoeld om de aspecten vindbaarheid, volgbaarheid en archiveerbaarheid voor dit onderdeel van het WSS te onderzoeken. Dit staat hieronder uitgewerkt in TABEL 14.

Work practices en processen afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid				
Activiteit	VDB[1]	VGB[2]	ARB[3]	Kort Commentaar
Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan;	X			Locatie voor vindbaarheid.
Het werksysteem keurt de aanvraag en kent ruimte toe of niet;	X			Locatie voor vindbaarheid.
Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte.	X			Locatie voor vindbaarheid.
De researcher verplaatst (meta)data op de geboden opslagruimte;	X			Op voorwaarde van geschikte logging vindbaar.
De researcher verwijdert data van de geboden opslagruimte;	X			Op voorwaarde van geschikte logging vindbaar.
De researcher bedenkt en gebruikt een heldere mappenstructuur op het werksysteem;	X			De structuur vereenvoudigt vinden.
Het werksysteem synchroniseert lokale onderzoeksdata en metadata met de centraal aangeboden opslagruimte;	X			Resultaat van decentrale handelingen wordt centraal opgeslagen.
De researcher slaat zijn (meta)data op op de geboden opslagruimte;	X			Locatie voor vindbaarheid.
De researcher wijzigt (meta)data op de geboden opslagruimte;		X		Op voorwaarde van geschikte logging volgbaar.
Het werksysteem voert versiebeheer uit over de (meta)data;		X		Helpt datatransformaties inzichtelijk te maken.
Het werksysteem levert de researcher mogelijkheden voor de aanmaak van metadata;		X		Metadata (datadocumentatie) verhoogt de volgbaarheid.
Het werksysteem voert backups uit op basis van de eisen van de researcher;		X		Backups geven tussentijdse versie en helpen volgbaarheid.
De researcher vraagt een PID aan bij het werksysteem voor een relevante datasets en bijbehorende metadata;			X	Een PID is nodig voor archiveerbaarheid.

Activiteit	VDB[1]	VGB[2]	ARB[3]	Kort Commentaar
Het werksysteem maakt een PID aan voor een dataset en bijbehorende metadata.			X	Een PID is nodig voor archiveerbaarheid.
De researcher vraagt toegang tot (delen van) de data aan voor derden;				Geen relatie met vind- of volgbaarheid.
Het werksysteem geeft derden toegang tot (delen van) de data;				Geen relatie met vind- of volgbaarheid.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid [3]ARB: Archiveerbaarheid				

Tabel 14, Work practices en processen afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid

2.7.5. Information binnen het werksysteem



Figuur 11, informatie voor het WSS

Deze paragraaf beschrijft de information die benodigd is om de work practices die zijn benoemd in 2.7.4 te kunnen uitvoeren. Om die reden wordt aan het eind van deze paragraaf niet meer gecontroleerd op vindbaarheid, volgbaarheid en archiveerbaarheid. De information-uitwisseling per activiteit staat beschreven in TABEL 15. De mogelijke communicatievormen worden beschreven in TABEL 16.

Informationuitwisseling per activiteit		
Activiteit	Informationuitwisseling tussen	Communi-catievorm[11]
Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan.	• Een researcher (customer) vraagt aan bij een technisch systeem in het werksysteem (userinterface). • De aanvraag gaat digitaal via DMP* naar een technisch systeem in het werksysteem. • De aanvraag wordt door een participant (bijvoorbeeld IT'er of onderzoek coördinator) binnen het werksysteem ingediend op basis van het DMP.	3, 9 en 7
Het werksysteem keurt de aanvraag en kent ruimte toe of niet.	• Een participant, bijvoorbeeld een afdelingshoofd, geeft op basis van het DMP toestemming via een userinterface aan het werksysteem. • Een technisch systeem binnen het werksysteem beslist op basis van de informatie in het CRIS** en/of in het DMP. • In beide gevallen geeft het werksysteem bericht aan de aanvrager (customer).	6, 9 en 3
Een researcher vraagt extra opslagruimte voor zijn onderzoeks(meta)data aan.	• Researcher (customer) vraagt rechtstreeks aan bij systeem in werksysteem (userinterface). • Systeem in werksysteem haalt aanvraag uit DMP. • Participant (bijvoorbeeld IT'er of onderzoek coördinator) haalt aanvraag uit DMP.	3, 9 en 7

[11] Zie Tabel 16

Activiteit	Informationuitwisseling tussen	Communi-catievorm[11]
Het werksysteem keurt de tussentijdse aanvraag en kent ruimte toe of niet.	• Een participant, bijvoorbeeld een afdelingshoofd, geeft op basis van het DMP toestemming via een userinterface aan het werksysteem. • Een technisch systeem binnen het werksysteem beslist op basis van de informatie in het CRIS en/of DMP. • Dat wordt via het userinterface teruggekoppeld aan de aanvrager (customer).	6, 9 en 3
De researcher bedenkt en gebruikt een heldere mappenstructuur.	• Een researcher maakt de mappenstructuur aan (userinterface). • De mappenstructuur wordt door een systeem in werksysteem uit het DMP gelezen (mappenstructuur wordt ook opgegeven in DMP).	3 en 9
De researcher slaat zijn (meta)data op op de geboden opslagruimte.	Tussen researcher (customer) en systeem in werksysteem (userinterface). (Of het gelukt is of niet, of er voldoende ruimte is of niet, wat de status van de handeling is.)	3
De researcher wijzigt (meta)data op de geboden opslagruimte.	Tussen researcher (customer) en systeem in werksysteem (userinterface). (Of het gelukt is of niet, of er voldoende ruimte is of niet, wat de status van de handeling is.)	3
De researcher verplaatst (meta)data op de geboden opslagruimte.	Tussen researcher (customer) en systeem in werksysteem (userinterface). (Of het gelukt is of niet, of er voldoende ruimte is of niet, wat de status van de handeling is.)	3
De researcher verwijdert (meta)data van de geboden opslagruimte.	Tussen researcher (customer) en systeem in werksysteem (userinterface). (Of het gelukt is of niet, of er voldoende ruimte is of niet, wat de status van de handeling is.)	3
De researcher vraagt toegang tot (delen van) de data aan voor derden.	Tussen researcher (customer) en systeem in werksysteem (userinterface) of tussen researcher en participant in werksysteem. (Wie waar welke toegang moet krijgen.)	3, 6
Het werksysteem geeft derden toegang tot (delen van) de data.	• Het werksysteem zal de aanvraag controleren in het DMP (via systeem in werksysteem of via een participant) en vervolgens al dan niet toestemming verlenen en de juiste autorisaties zetten. • Dat wordt via het userinterface teruggekoppeld aan de aanvrager (customer).	6, 9 en 3
Het werksysteem synchroniseert lokale onderzoeksdata en metadata met de centraal aangeboden opslagruimte.	Het systeem buiten het werksysteem constateert een wijziging in de data en zet die wijziging door op het systeem binnen het werksysteem (datasynchronisatie).	9
Het werksysteem voert versiebeheer uit over de (meta)data.	• Informatie hierover komt uit het DMP al dan niet uitgelezen via een systeem in het werksysteem of via een participant. • Het kan dat een onderzoeker (customer) handmatig versienummers toekent aan datasets en/of documentatie. • Het kan ook dat gewijzigde versies automatisch worden gedetecteerd door een versiebeheersysteem en dat deze de versienummering verzorgt. • In alle gevallen zal het resultaat voor de customer zichtbaar zijn via een userinterface.	7, 9, 3 en 8
Het werksysteem levert de researcher mogelijkheden voor de aanmaak van metadata.	• Informatie hierover komt uit het DMP al dan niet uitgelezen via een systeem in het werksysteem of via een participant. • Een onderzoeker (customer) kan handmatig metadata aanmaken (bijvoorbeeld in de vorm van een document). • Het kan ook metadata geautomatiseerd wordt aangemaakt door een systeem in het werksysteem. • In alle gevallen zal het resultaat voor de customer zichtbaar zijn via een userinterface	7, 9, 3 en 8

Activiteit	Informationuitwisseling tussen	Communi-catievorm[11]
Het werksysteem voert backups uit op basis van de eisen van de researcher.	• De eisen van de researcher Informatie hierover komen uit het DMP al dan niet uitgelezen via een systeem in het werksysteem of via een participant. • Het resultaat zal aan de customer getoond worden via een userinterface.	7, 9 en 3
De researcher vraagt een PID aan bij het werksysteem voor een relevante dataset en bijbehorende metadata.	Aan het eind van de tijdens fase, vraagt de researcher (customer) een PID aan bij een systeem in werksysteem of aan een participant voor de relevante datasets en metadata.	3, 6
Het werksysteem maakt een PID aan voor een dataset en bijbehorende metadata.	Het werksysteem maakt een PID aan (eventueel op basis van informatie uit het archiveringssysteem). Het resultaat zal zichtbaar zijn via een userinterface.	8, 3
Opmerkingen: *DMP is het data management plan dat voor een onderzoek moet worden aangemaakt. **Het CRIS is een informatiesysteem dat voldoet aan de eisen dat alle stakeholders er zicht op hebben en gedeelde verantwoordelijkheid kunnen dragen. Het is echter in de hier genoemde zin ondersteunend aan managing active data en wordt om die reden als systeem buiten het werksysteem beschouwd. In de bovenstaande work practices, vindt er geen directe communicatie plaats tussen de customer en de participant of tussen participants onderling. Zij communiceren via het werksysteem.		

Tabel 15, informationuitwisseling per activiteit

De verschillende vormen van communicatieuitwisseling staan op de volgende pagina in TABEL 16. Hieruit blijkt dat de belangrijkste informatiebronnen zijn (met tussen haakjes een summier voorbeeld):

- Informatieuitwisseling tussen customer en participant (een aanvraag voor onderzoeksopslag) ingevoerd in het userinterface (2->3, 6);
- Informatieuitwisseling tussen een systeem binnen het werksysteem en de customer (informatie tonen op userinterface) (3).
- Informatieuitwisseling tussen een systeem binnen het werksysteem en een participant (een onderzoek coördinator geeft toestemming om opslag voor een onderzoek toe te kennen) (6).
- Informatieuitwisseling tussen een participant en een systeem buiten het werksysteem (bijvoorbeeld een aanvraag controleren op het DMP) (7).
- Informatieuitwisseling tussen systemen binnen het werksysteem (nadat het opslagsysteem een schrijfactie doorgeeft, gaat een ander systeem metadata toevoegen) (8).
- Informatieuitwisseling tussen een systeem binnen het werksysteem en een systeem er buiten (het controleren van een DMP voor het geautomatiseerd toekennen van een autorisatie) (9).

Vormen van communicatieuitwisseling van het werksysteem				
	Customer	Participant	Systeem in werksysteem	Systeem buiten werksysteem[12]
Customer	1	2	3	4
Participant	2	5	6	7
Systeem in Werksysteem	3	6	8	9
Systeem buiten werksysteem	4	7	9	10

verklaring:

1. Customers hebben bij de voorgestelde work practices geen interactie met elkaar.
2. Een customer deelt information met een participant. Dat gaat indirect via een userinterface. Dat zal gaat via een userinterface. Dat betekent dat de customer in dit geval met een systeem in het werksysteem communiceert (3) en de participant met een systeem in het werksysteem communiceert (6).
3. Een customer deelt information met een technisch (deel)systeem, dit kan bijvoorbeeld een aanvraag van een onderzoeker om opslagruimte zijn als dit proces is geautomatiseerd.
4. Customers hebben bij de voorgestelde work practices geen interactie met systemen buiten het werksysteem.
5. Participants hebben bij de voorgestelde work practices geen interactie met elkaar.
6. Een participant deelt information met een systeem in het werksysteem (hij geeft bijvoorbeeld een autorisatie).
7. Een participant haalt benodigde information voor werkzaamheden uit een extern systeem zoals het CRIS of een digitaal DMP.
8. Een systeem binnen het werksysteem kan reageren op input van een ander of hetzelfde systeem binnen een werksysteem.
9. Een systeem binnen het werksysteem haalt benodigde information voor werkzaamheden uit een extern systeem zoals het CRIS of een digitaal DMP.
10. Systemen buiten het werksysteem hebben bij de voorgestelde work practices geen interactie met elkaar.

Tabel 16, informationuitwisseling in het werksysteem

2.7.6. Ingevuld worksystem snapshot

Op basis van de resultaten beschreven in de tabellen TABEL 12, TABEL 13 en TABEL 14 worden er in BIJLAGE SAMENVOEGING VAN DE RESULTATEN VOOR PRODUCTS AND SERVICES EN TECHNOLOGY T.A.V. VINDBAARHEID, VOLGBAARHEID EN ARCHIVEERBAARHEID zestien uitspraken over het te bouwen werksysteem, waarbij de aspecten vindbaarheid, volgbaarheid en/of archiveerbaarheid een rol spelen, benoemd. Het zijn veertien uitspraken over 'products and services' en twee technology bevindingen. Ze worden gebruikt om het WSS te vullen. In TABEL 17 staat de eerste versie van het ingevulde WSS.

[12] Systemen buiten het werksysteem waarmee communicatie plaatsvindt zijn van belang voor het resultaat van het werksysteem. Als zodanig maken ze onderdeel uit van de infrastructuur.

<table>
<tr><th colspan="3">Work system snapshot voor managing active data die vindbaar, volgbaar en archiveerbaar zijn</th></tr>
<tr><th>Customers</th><th colspan="2">Products & Services</th></tr>
<tr><td>• Research communities (onderzoeksgemeenschappen, vaak ook over instituten heen);
• Researchers (onderzoekers en master studenten);
• Research coordinators (onderzoekscoördinatoren);
• Faculties and departments (faculteiten en afdelingen van faculteiten).
• Computational stakeholders.</td><td colspan="2">• Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.
• Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders.
• Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.
• Het werksysteem beschrijft (meta)data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR principes volgen.
• Het werksysteem maakt metadata voorkeur geautomatiseerd aan.
• Het werksysteem verzekert dat restoremogelijkheden gedurende het onderzoek beschikbaar zijn.
• Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.
• Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen.
• Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.
• Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering).
• Het werksysteem stelt (meta)data met een duidelijke en toegankelijke dataovereenkomst beschikbaar.
• Het werksysteem kan unieke identifiers aanmaken voor relevante onderzoeksdata (Persistent Identifiers, PID). De wetenschapper bepaalt hierbij wat relevante onderzoeksdata zijn.
• Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.
• Het werksysteem heeft een userinterface en of API om informatie te tonen, work practices te volgen, logs te lezen en (zoek)opdrachten te geven.</td></tr>
<tr><th colspan="3">Work practices</th></tr>
<tr><td colspan="3">• Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan;
• Het werksysteem keurt de aanvraag en kent ruimte toe of niet;
• Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte.
• De researcher slaat zijn (meta)data op de geboden opslagruimte;
• De researcher wijzigt (meta)data op de geboden opslagruimte;
• De researcher verplaatst (meta)data op de geboden opslagruimte;
• De researcher verwijdert (meta)data van de geboden opslagruimte;
• De researcher bedenkt en gebruikt een heldere mappenstructuur;
• Het werksysteem synchroniseert lokale onderzoeksdata en metadata met de centraal aangeboden opslagruimte;
• Het werksysteem voert versiebeheer uit over de (meta)data;
• Het werksysteem levert de researcher mogelijkheden voor de aanmaak van metadata;
• De researcher vraagt een PID aan bij het werksysteem voor een relevante datasets en bijbehorende metadata;
• Het werksysteem maakt een PID aan voor een dataset en bijbehorende metadata.</td></tr>
<tr><th>participants</th><th>Information</th><th>Technologies</th></tr>
<tr><td>• Research communities (onderzoeksgemeenschappen, vaak ook over instituten heen);
• Researchers (onderzoekers en master studenten);
• Research coordinators (onderzoekscoördinatoren);
• Faculties and departments (faculteiten en afdelingen van faculteiten).
• Research libraries;
• Campus IT services;
• Infrastructure providers (mogelijk samenwerkende aanbieders van technische infrastructuur voor RDM diensten);
• Federated groups (samenwerkingsverbanden van (onderdelen van) verschillende instituten om een (deel van een) RDM dienst aan te bieden);
• Offices of research
• Computational stakeholders.</td><td>• Informationuitwisseling tussen customer en participant (een aanvraag voor onderzoeksopslag) ingevoerd in het userinterface;
• Informationuitwisseling tussen een systeem binnen het werksysteem en de customer (information tonen op userinterface);
• Informationuitwisseling tussen een systeem binnen het werksysteem en een participant (een onderzoek coördinator geeft toestemming om opslag voor een onderzoek toe te kennen);
• Informationuitwisseling tussen een participant en een systeem buiten het werksysteem (bijvoorbeeld een aanvraag controleren op het DMP);
• Informationuitwisseling tussen systemen binnen het werksysteem (nadat het opslagsysteem een schrijfactie doorgeeft, gaat een ander systeem metadata toevoegen);
• Informationuitwisseling tussen een systeem binnen het werksysteem en een systeem er buiten (het controleren van een DMP voor het geautomatiseerd toekennen van een autorisatie).</td><td>• Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface.
• Het werksysteem biedt opslag voor de RDM infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd).</td></tr>
</table>

Tabel 17, Work system snapshot voor managing active data die vindbaar, volgbaar en archiveerbaar zijn

2.8. Conclusies en aanbevelingen

De conclusies worden opgesplitst in een inhoudelijk deel en een deel waarin op de gebruikte methode wordt ingegaan. Er wordt afgesloten met de aanbevelingen voor het vervolgonderzoek.

2.8.1. Inhoudelijke conclusies

Bij RDM is het beheer van alle onderzoeksdata gedurende hun gehele levensduur van belang. De meeste literatuur legt de nadruk op langdurige dataopslag na afloop van een onderzoek. In dit onderzoek worden alleen de data die gedurende het lopende onderzoek gebruikt worden, beschouwd. Het WSS (TABEL 17) geeft een overzicht van wat specifiek voor managing active data is gevonden met betrekking tot vindbaarheid, volgbaarheid en archiveerbaarheid. Het hart van het WSS wordt gevormd door de products and services en de technology. Ze zijn gebaseerd op de opgebouwde theorie. Work practices en information zijn er van afgeleid en customers en participants volgen uit de literatuur. De products and services en technology zijn:

Vindbaar

- Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.
- Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders.
- Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.

Vindbaar en volgbaar

- Het werksysteem beschrijft (meta)data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR principes volgen.
- Het werksysteem ondersteunt het aanmaken van metadata (bij voorkeur geautomatiseerd).
- Het werksysteem verzekert dat restoremogelijkheden gedurende het onderzoek beschikbaar zijn.
- Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.

Volgbaar

- Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen.
- Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.

Volgbaar en archiveerbaar

- Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering).

Archiveerbaar

- Het werksysteem stelt (meta)data met een duidelijke en toegankelijke dataovereenkomst beschikbaar.

- Het werksysteem kan unieke identifiers aanmaken voor relevante onderzoeksdata (Persistent Identifiers, PID). De wetenschapper bepaalt hierbij wat relevante onderzoeksdata zijn.

Vindbaar, volgbaar en archiveerbaar

- Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.
- Het werksysteem heeft een userinterface en API om informatie te tonen, work practices te volgen, logs te lezen en (zoek)opdrachten te geven.

Technology:

- Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface.
- Het werksysteem biedt opslag voor de RDM infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd).

2.8.2. Conclusies met betrekking tot de gebruikte methode

Voor RDM is voldoende informatie in de literatuur te vinden om een WSS te vullen. Door het te toetsen aan de aspecten vindbaarheid, volgbaarheid en archiveerbaarheid uit de probleemstelling en het te beperken tot de 'tijdens' fase, is het ook te specificeren tot het 'managing active data' onderdeel van RDM. Op het eerste gezicht, komt daar een rijk gevuld WSS uit. In de empirische fase zal kan dit een eerste aanzet zijn voor een case organisatie.

Er wordt op één belangrijk punt van de WSM afgeweken:
In het WSS zijn de eisen aan de te bouwen infrastructure gecombineerd met de technology (hulp)middelen in het werksysteem, aangezien die laatste van infrastructurele aard zijn en er geen bestaande RDM infrastructuur is. In de empirische fase zal geëvalueerd moeten worden of dit tot relevante resultaten leidt en geen relevante resultaten blokkeert. Na de empirische fase kunnen ook conclusies getrokken worden over het gebruik van DSR en over de combinatie van DSR en de WSM.

2.8.3. Doel en scope van het vervolgonderzoek

In de introductie werd al gerefereerd aan het feit dat het probleem van her en der opgeslagen onderzoeksdata bij de case organisatie een eerste aanzet was voor dit onderzoek. De meeste gebruikte literatuur is buitenlands, maar wel toegespitst op universiteiten en ook hier lijkt het probleem van verspreide opslag zich voor te doen. Door een case organisatie te gebruiken, kan er gecontroleerd worden of de bevindingen uit de literatuur ook hier geldig zijn en zo ja, of er aanvullingen zijn op de bevindingen uit de literatuur. De mogelijkheid dat de praktijk op de case organisatie anders is, dient natuurlijk ook open gehouden te worden. In de empirische fase kan daardoor de praktische relevantie van het ontwerp (WSS) versterkt worden en gecontroleerd worden of bevindingen uit de literatuur generaliseerbaar zijn naar de case organisatie. De mate waarin die generaliseerbaarheid optreedt, zou een indicatie kunnen zijn voor de praktische bruikbaarheid van het ontwerp op andere universiteiten.

Het WSS uit de literatuur is relatief groot. Er staan 50 bevindingen in. Voor dit onderzoek gaat het te ver om alle bevindingen in de empirische fase verder te onderzoeken. Om deze reden is gekozen de aandacht te richten op products and services en technology in de empirische fase. De verwachting is wel dat customers en participants impliciet aan bod zullen komen in de interviews, zodat ze in het eindresultaat benoemd kunnen worden zoals ze bij de case organisatie zijn ingevuld. Om de scope van het onderzoek verder beperkt te houden, wordt besloten om af te zien van de confrontatie van het WSS met de architectuurprincipes. Hiermee wordt nog steeds voldaan aan de voorwaarden voor DSR, waarbij een design artifact in de vorm van een model wordt opgeleverd.
In de literatuurstudie zijn de work practices van de products and services afgeleid. De informationuitwisseling is hier weer van afgeleid. Technology en infrastructure zijn samengevoegd in de literatuur omdat ze sterk op elkaar leken aan te sluiten (met name opslag). Er is nog geen reden om dat aan te passen.

3. Methodologie

In de empirische fase wordt na bijstelling van de scope en bepaling van het doel van deze fase de volgende probleemstelling gebruikt:

Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op het beheer van actieve onderzoeksdata, binnen de case organisatie, zodat deze data vindbaar, volgbaar en archiveerbaar zijn?

De bijbehorende deelvragen voor deze fase zijn:

1. Wat verstaat men bij de case organisatie onder actieve onderzoeksdata?
2. Hoe maakt men data vindbaar?
3. Hoe maakt men data volgbaar?
4. Hoe maakt men data archiveerbaar?
5. Hoe wordt invulling gegeven aan het resulterende WSS?
6. Hoe beoordelen de experts de producten en diensten uit het WSS?

In dit hoofdstuk wordt ingegaan op de gekozen onderzoeksmethode om deze vragen te beantwoorden.

3.1. Conceptueel ontwerp: keuze van onderzoeksmethode

Het eerste doel van de empirische fase is achterhalen in hoeverre het WSS uit de literatuur ook past bij de case organisatie. De bedoeling is te begrijpen hoe de onderzoekers bij de case organisatie werken en waarom ze het zo doen. Op basis daarvan kan het WSS verbeterd worden. Ten slotte worden experts gevraagd om hun op kennis en ervaring gebaseerde oordeel te geven over de products and services in het WSS.

Het WSS geeft een momentopname. Het weerspiegelt de werkwijze en de voorkeuren van de organisatie t.a.v. bepaalde werkzaamheden op een bepaald moment. Het invulling geven aan het WSS voor de case organisatie vergt begrip van hoe de organisatie werkt en waarom. De theorievorming is daarmee inductief en vraagt om kwalitatieve methodes (Bryman 2012; Mark Saunders, Philip Lewis 2015).[13]

Het WSS wordt niet gebruikt om ontwikkelingen in de loop van de tijd vast te stellen. Het gaat meer om doorsnede onderzoek dan om longitudinaal onderzoek (Bryman 2012; Mark Saunders, Philip Lewis 2015).

Bryman (2012) geeft in tabel 3.1. een overzicht van verschillende benaderingen van onderzoek en welke methodes daarbij gebruikt kunnen worden. Voor kwalitatief doorsnede onderzoek geldt:

- Kwalitatieve interviews of focusgroepen;
- Kwalitatief documentonderzoek uit een bepaalde periode.

Er is bij de case organisatie nauwelijks tot geen documentatie aanwezig om documentonderzoek over RDM in de actieve fase op te kunnen doen.

[13] Bij voldoende zekerheid over de resultaten in de literatuur, zou men de verschillende concepten kunnen verifiëren in een enquête. In dit onderzoek wordt het WSS specifiek gemaakt en verdiept voor de case organisatie.

Er is dus een keuze tussen focusgroepen of interviews, of allebei.

Tremblay, Hevner, and Berndt (2008) beschrijven voor het gebruik van focusgroepen in DSR, zowel de voordelen als de beperkingen. Zij raden het gebruik van focusgroepen aan in de ontwerpfase en tijdens de veldtest (de laatste is geen onderdeel van dit onderzoek). Een aantal relevante beperkingen dat wordt genoemd:

1. Het is ingewikkeld om een groep met de juiste kennis bij elkaar te krijgen;
2. In de ontwerpfase zou men net zo lang door moeten gaan met focusgroepen tot er geen nieuwe inzichten meer komen. Dit maakt het aantal sessies van de focusgroepen vooraf lastig in te schatten;
3. Een deelnemer met een sterke mening zou de anderen kunnen overstemmen;
4. Het succes van de focusgroep staat of valt bij de vaardigheden van de moderator om de groep optimaal te laten samenwerken en de resultaten niet te sturen of laten sturen.

Voor punt één en twee: Het blijkt in de praktijk inderdaad lastig om verschillende stakeholders (bijvoorbeeld data stewards en ICT ondersteuners) op hetzelfde moment bij elkaar te krijgen. De agenda's zijn vol. Een aantal keer achter elkaar wat inplannen, maakt het nog lastiger.
Voor punt drie en vier: Er is geen ervaren moderator. Niet optimale samenwerking en het beperken van het overstemmen van elkaar is een reële mogelijkheid.

Er wordt gekozen voor semigestructureerde interviews en theorievorming met grounded theory. Een expert panel wordt gevraagd om te valideren of het ontwerp in de praktijk kan voldoen.

3.2. Technisch ontwerp: uitwerking van de methode

Semigestructureerde interviews met grounded theory

Om het WSS specifiek voor de case organisatie te maken, is er informatie nodig van de mensen die zich bezig houden met RDM van actieve onderzoeksdata. Om te bepalen wie de gebruikers zijn, kan de lijst met stakeholders, zoals geformuleerd in de literatuurstudie, gebruikt worden nadat hij is gespecificeerd naar de case organisatie (SELECTIE GEÏNTERVIEWDEN). Specifiek voor de case organisatie zijn dat:

- Afdelingshoofden;
- Principle Investigators (PI's)[14];
- Onderzoekers;
- Data Stewards;
- IT ondersteuners (in de afdeling/sectie).

Deze personen hebben allemaal te maken met management van onderzoeksdata, hetzij in hun dagelijkse onderzoekspraktijk, hetzij als adviseurs of ondersteuners t.a.v. het omgaan met en het beheer van onderzoeksdata.

[14] De PI's zijn de onderzoekscoördinatoren/hoofdonderzoekers.

Hoe de beoogde interviewkandidaten zijn geselecteerd en benaderd staat in SELECTIE GEÏNTERVIEWDEN en BESCHRIJVING GEÏNTERVIEWDEN. Dit leidt tot TABEL 18 waarin staat aangegeven welke rol een functie kan aannemen en welke rol die speelt in een onderzoek.

Lijst met rollen geïnterviewden en hun invloed binnen een onderzoek						
Rol	Rol in onderzoek	Benaderd via	Data Champion?	Bepaalt RDM Richtlijnen	Adviseert RDM Richtlijnen	Onder-steunt onderzoek
HL	PI	ICT	Ja/Nee	Afd/grp	Afd/groep	Afd/groep
UHD	PI	HPC	Ja/Nee	Afd/grp	Grp	Grp
UD	PI	HPC	Ja/Nee	Afd/grp	Grp	Grp
Postdoc	Onderzoeker	DS	Ja/Nee	/	Grp	Grp
DS	DataSteward	DS	Ja/Nee	Fac	Fac/Afd/groep	Fac/Afd/groep
Support	/	DS	Ja/Nee	/	/	Afd/groep

Opmerkingen: Voor de vergelijkbaarheid met andere Nederlandse universiteiten staan de rollen in het Nederlands afgekort, hoewel de case organisatie Engelse benamingen voert:
HL = hoogleraar => Professor
UHD = Universitair HoofdDocent => Associate Professor
UD Universitair Docent => Assistant Professor
HPC is high performance computing, hier wordt de universiteit brede groep bedoeld, die HPC coördineert.
PI = Principle investigator
DS = Data Steward
Fac = faculteit
Afd = Afdeling binnen een faculteit
Grp = Onderzoeksgroep

De tabel geeft aan waar personen mogelijk invloed kunnen uitoefenen en hoe. Daar waar PI genoemd staat, betekent dat de persoon als PI kan optreden, dat wil niet zeggen dat dat ook altijd zo is.

Professors, Associate Professors en Assistant Professors doen of deden allen onderzoek en/of geven leiding aan mensen die onderzoek doen. Zij kunnen allemaal als PI optreden. De professor is ook leerstoelhouder of afdelingshoofd en kan in die zin RDM beleid bepalen voor iedereen in de afdeling.

Een PI is altijd ook een onderzoeker. Een PI kan beleid bepalen rondom RDM in de onderzoeksgroep, maar kan er ook voor kiezen te adviseren en/of te ondersteunen. Dat is redelijk vrij.

De professor zal meer aan de bepalende kant staan van de RDM richtlijnen in de onderzoeksgroep/afdeling, een Associate en Assistant Professor meer aan de adviserende en ondersteunende kant.

De data stewards schrijven RDM beleid voor de case organisatie en de faculteiten. Dit is niet bindend en de onderzoekers kunnen daar van afwijken.

Tabel 18, Lijst met rollen geïnterviewden en hun invloed binnen een onderzoek

In eerste instantie werden er twee interviews[15] gehouden met twee onderzoekers, waarbij een deel gestructureerd was. Het WSS werd aan ze voorgelegd. Per onderdeel werd gevraagd of ze zich er in konden vinden en of ze eventueel aanvullingen hadden (zie INITIEEL SCRIPT). De interviewopzet bleek niet te passen. Beide wetenschappers hadden geen interesse om aandacht aan het WSS te besteden. Deze manier van interviewen bleek ook veel voorbereidingstijd bij de geïnterviewden te kosten (beiden hadden de vooraf gestuurde stukken niet gelezen)[16].

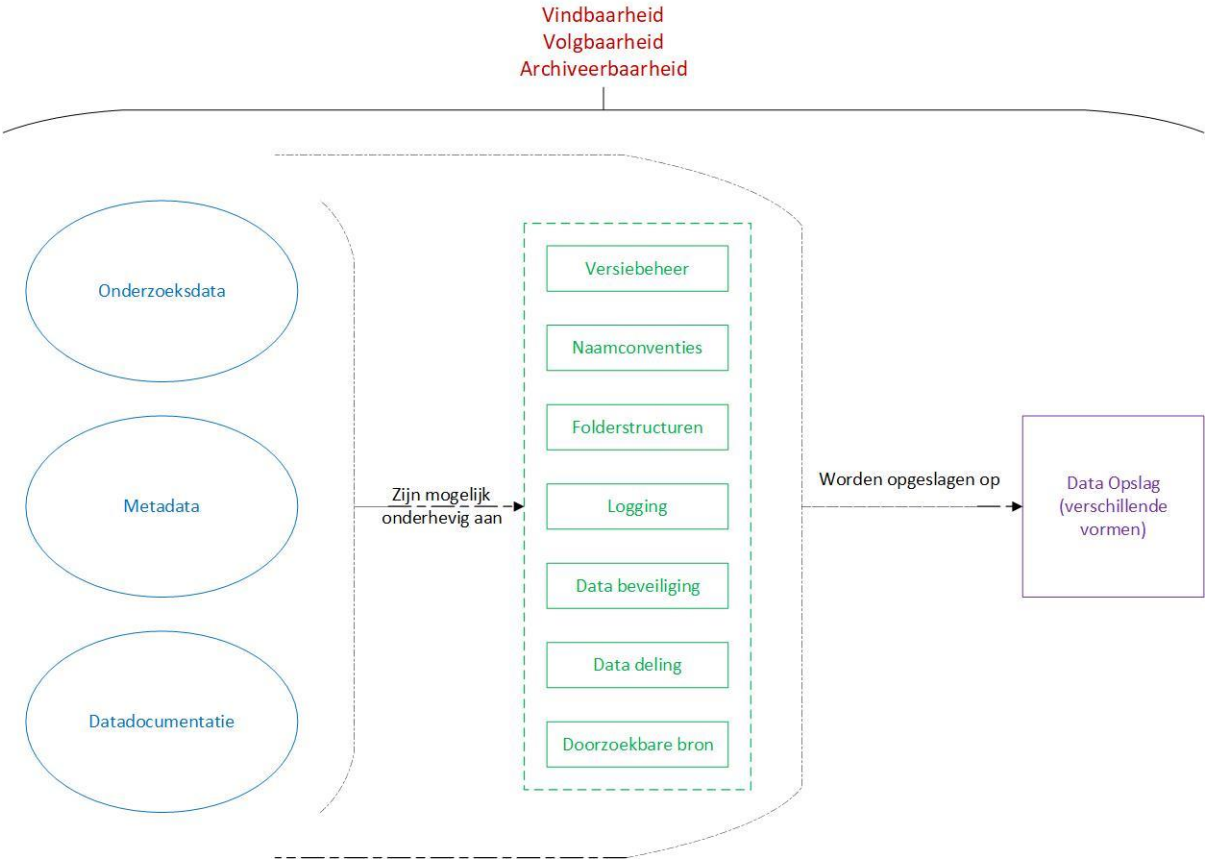
[15] Persoonlijke communicaties op 3-12-2018 en 19-12-2018

[16] De persoonlijke communicatie van 3-12-2018 is uiteindelijk wel opgenomen in de resultatenanalyse omdat er voldoende input was geleverd voor het onderwerp. Het interview van 19-12-2018 is uiteindelijk niet opgenomen vanwege gebrek aan input.

De aanpak werd daarom gewijzigd. In de literatuur werden vooral products and services en enigszins technology gevonden. Work practices en information in het WSS werden er van afgeleid. Voor het empirische deel werd daarom een zelfde aanpak gekozen: De focus werd gelegd op de products and services zoals beschreven in de literatuurstudie en de technology.

Deze werden vertaald in concepten (zie UITWERKING TOTSTANDKOMING CONCEPTEN en FIGUUR 12). Per concept (dataconcepten ovaal in de tekening, overige concepten rechthoekig) werd een aantal vragen voorbereid. Met vindbaarheid, volgbaarheid en archiveerbaarheid zijn dat veertien onderwerpen van gesprek per interview. Het script voor het interview bestond uit deze vragen (UITEINDELIJK SCRIPT). De uitkomsten van de literatuurstudie werden niet meer vooraf gezonden.

Deze semigestructureerde manier van interviewen zou het gemakkelijker moeten maken om in gesprek te raken met de geïnterviewde en hem/haar de ruimte moeten geven om te kunnen vertellen hoe hij/zij RDM aanpakt in zijn/haar onderzoek. Op deze manier werd minder expliciet gezocht naar verificatie van de uitkomsten van de literatuurstudie en werd vooral gevraagd naar products and services en technology van het WSS. De information, stakeholders en de work practices kwamen impliciet aan bod in de interviews.



Figuur 12, conceptuele voorstelling van de bevindingen uit de literatuurstudie

Belangrijkste voordelen van deze aanpak zijn:

- Vrijheid voor de onderzoeker om te vertellen hoe de dagelijkse praktijk rondom RDM er uit ziet;

- De mogelijkheid om te achterhalen waarom iemand op een bepaalde manier werkt;
- Ruimte voor individuele inbreng;
- De mogelijkheid tot theorievorming;
- Een foutje door onervarenheid in een interview, is waarschijnlijk te corrigeren (in hetzelfde interview of in de volgende interviews).

Nadelen zijn:

- Als niet het WSS uit de literatuurstudie als referentie wordt gebruikt, maar de dagelijkse praktijk van de geïnterviewden, dan zou dat kunnen betekenen dat beiden zo sterk afwijken dat de volgende versie van het WSS geen evolutie is, maar een nieuw model. Dat is dan wel specifiek voor de case organisatie, maar generaliseert waarschijnlijk minder goed naar andere (westerse) universiteiten;
- Door te interviewen, kun je goed naar de overwegingen van een persoon vragen, maar ontstaat er geen groepsdiscussie waarin de mening van de één de gedachte van de ander helpt vormen.

Expert review

Uiteindelijk worden de bevindingen van de products and services voorgelegd aan een expertpanel. Dit panel werd samengesteld met een proces architect en een infrastructuur architect, zodat de producten en diensten en de infrastructuur in de structuur van de instelling kunnen passen, alsmede een onderzoeksadviseur bij de ICT afdeling Innovatie, zodat de toepassing in de dagelijkse praktijk van wetenschappers kan passen. (zie BIJLAGE SELECTIE EXPERT PANEL). Er is gebruik gemaakt van de Delphi Methode (Yousuf 2007). Drie experts kregen onafhankelijk van elkaar de products and services zoals beschreven na de interviews, uitleg over het proces hoe tot deze products and services gekomen was en wat extra uitleg over gebruikte begrippen (zie BIJLAGE CORRESPONDENTIE EN BIJLAGEN BIJ DE EXPERT REVIEW). Vervolgens werd ze gevraagd elke product/dienst te scoren op een Likert schaal: kan ik me helemaal niet in vinden, kan ik me niet in vinden, sta ik neutraal in, kan ik me in vinden, kan ik me helemaal in vinden, geen oordeel. Het verschil tussen sta ik neutraal in en geen oordeel is:

- Sta ik neutraal in: Iemand is niet voor of tegen;
- Geen oordeel: iemand geeft aan dat hij/zij zich niet geschikt acht om hier een gewogen oordeel te geven.

Bovendien werd verzocht de scores van commentaar te voorzien. Na elke ronde werd een voorstel gedaan voor een aangepaste formulering van de products and services op basis van het geleverde commentaar. Bovendien kon men tussen de rondes in elkaars commentaar lezen. De experts werden na elke ronde verzocht te scoren en commentaar te geven. Na de eerste en tweede ronde werd bovendien verduidelijkende informatie gegeven. Men kwam tot een eindscore in de derde ronde. Alles verliep schriftelijk en anoniem. De experts hebben elkaars namen nooit gezien en wisten niet van elkaar dat ze in de reviewgroep zaten.

3.3. Gegevensanalyse

Er zijn 14 interviews afgenomen met personen met rollen zoals beschreven in (TABEL 18). De gehele lijst staat beschreven in (BESCHRIJVING GEÏNTERVIEWDEN). De interviews werden (met toestemming van

de geïnterviewde) allemaal opgenomen. De opnames werden getranscribeerd met behulp van Amberscript[17]. Vervolgens werden ze met de hand nagelopen en verder gecorrigeerd.

Voor de gegevensanalyse is gebruik gemaakt van een CAQDAS pakket (Computer Assisted Qualitative Data Analysis Software). Hiervoor werd Atlas.ti[18] versie 8 gebruikt. Alle transcripties werden in het CAQDAS pakket geladen. Hierna werden ze gecodeerd. Het gevolgde proces is gebaseerd op Silver and Woolf (2018) en Friese (2017):

1. Getranscribeerde interviews laden in CAQDAS pakket;
2. Drie interviews uitzoeken die ieder een breed spectrum aan onderwerpen behandelden en die één voor één coderen;
    a. Belangrijke uitspraken oormerken (een quotation maken);
    b. De quotation van één of meer codes voorzien (bijvoorbeeld research data en metadata).
3. De codes van de drie interviews met elkaar vergelijken en waar nodig aanpassen:
    a. Indien nodig codes samenvoegen;
    b. Indien nodig codes verwijderen en een nieuwe toevoegen;
    c. Indien nodig codes splitsen, zoals voor meer detail informatie (bijvoorbeeld research data wordt research data definitie, research data actief, research data herkomst, een categorie met subcategorieën aan quotations).
4. De overige interviews coderen met de codeset uit punt 3.
5. Per code alle quotations onder elkaar zetten en controleren of de code valide is:
    a. Zo ja, dan kan de code blijven staan;
    b. Zo nee, dan wordt de code herzien:
        i. Andere (preciezere) beschrijving en naam voor de (sub)code;
        ii. Alle quotations uit de code categorie nog eens indelen in nieuwe subcategorieën.
        iii. Mogelijk kleine subcategorieën toevoegen aan een andere subcategorie door deze laatste een iets bredere definitie te geven.
6. De mate van verwantschap[19] tussen de subcategorieën bepalen (bijvoorbeeld het kunnen vinden van data met het hanteren van een naamconventie).
7. Met de literatuurstudie er bij, theorie opbouwen uit de gevonden codes en mogelijke onderlinge verwantschap van codes.

Met de uitkomsten hiervan, kan het WSS zo aangepast worden dat het past binnen de case organisatie, mogelijk kunnen ook begrippen gevonden in de literatuur, verrijkt worden met hetgeen in de interviews wordt opgehaald. De products and services worden vervolgens voorgelegd aan het expert panel (gegevens analyse volgens de Delphi methode staat in de vorige paragraaf).

[17] www.amberscript.com

[18] atlasti.com

[19] Verwantschap geeft een mogelijke relatie tussen twee codes aan (vergelijkbaar met een correlatie). Verwantschappen worden tot uitdrukking gebracht in de c-coefficient. In hoeverre ze relevant zijn, moet worden onderzocht per relatie (zie ook BIJLAGE UITLEG CO-OCCURRENCE/VERWANTSCHAP)

3.4. Reflectie t.a.v. validiteit, betrouwbaarheid en ethische aspecten

Gebruik maken van een niet gestandaardiseerde onderzoeksmethode heeft gevolgen voor de betrouwbaarheid, validiteit en generaliseerbaarheid van de resultaten.

3.4.1. Interne Validiteit

Intern valide onderzoek betekent dat de resultaten ook de werkelijkheid weerspiegelen. Interviews met mensen die de ruimte krijgen om te vertellen hoe ze werken en waarom ze bepaalde keuzes maken, is subjectief. Door te werken vanuit de concepten uit de literatuurstudie en te vragen naar die onderdelen van RDM, worden in de interviews i.i.g. dezelfde onderwerpen behandeld. De context zal voor elke geïnterviewde (minimaal) net even anders zijn. De verwachting is dat de interne validiteit niet hoog is. Als het op te leveren constructen (WSS) is uitgewerkt, zou verdere validatie mogelijk zijn met kwantitatief onderzoek (bijvoorbeeld een enquête).

3.4.2. Externe validiteit (generaliseerbaarheid)

De resultaten van de literatuurstudie zijn gebaseerd op veelal peer gereviewde artikelen. De resultaten daarvan hebben daardoor een zekere mate van generaliseerbaarheid. Als de resultaten bij de case organisatie in dezelfde lijn zijn, dan blijft die generaliseerbaarheid behouden en wordt zelfs met een extra geval bekrachtigd. Als de resultaten van de case organisatie bestaande uitkomsten aanvullen, kan dat een verrijking van de uitkomsten betekenen, maar of het ook weer terug generaliseerbaar is, zal moeten worden onderzocht. Als de uitkomsten voor de case organisatie sterk afwijken, past de case organisatie kennelijk niet in het model van de literatuurstudie en is de kans op generaliseerbaarheid van de resultaten waarschijnlijk laag.

3.4.3. Betrouwbaarheid

(Mark Saunders, Philip Lewis, p233, 2015) citeren (Marshall and Rossman 2006): “Resultaten die zijn afgeleid door het gebruik van niet-gestandaardiseerde onderzoeksmethoden hoeven niet noodzakelijkerwijs herhaalbaar te zijn, omdat ze de werkelijkheid weerspiegelen op het moment dat ze zijn verzameld in een situatie die aan veranderingen onderhevig kan zijn.” Zij stellen ook: “De waarde van het gebruik van niet-gestandaardiseerde interviews ligt juist in de flexibiliteit die je kunt gebruiken om de complexiteit van het onderwerp te onderzoeken.”

Het zou daarom niet realistisch zijn om te verwachten dat een andere onderzoeker dezelfde resultaten zou behalen (Mark Saunders, Philip Lewis 2015).

Ook het persoonlijke aspect van de codering maakt de betrouwbaarheid van het onderzoek lager (benoemd in GEGEVENSANALYSE).

De betrouwbaarheid zal niet hoog zijn, maar wordt zo sterk mogelijk gemaakt door:

- Het proces van de selectie van de interviewkandidaten goed te beschrijven;
- De onderwerpen (concepten) van de vragen goed te beschrijven;
- Transcripties te maken, zodat de interviews nagelezen kunnen worden;
- Een goed gedocumenteerd codeboek.

3.4.4. Ethiek

(Bryman 2012) gebruikt de vier belangrijkste gebieden zoals beschreven door (Diener and Crandall 1978) voor ethische principes:

Breng geen schade toe aan deelnemers (aan het onderzoek). Dit betekent dat:

- De deelname aan het onderzoek geen fysieke of mentale schade mag toebrengen.
    - Gesprekken worden gevoerd op locatie naar keuze van de deelnemer;
    - Er worden geen persoonlijke vragen gesteld, er wordt alleen naar werkwijze en meningen gevraagd uit de dagelijkse praktijk;
    - Men heeft de vrijheid geen antwoord te geven.
- Uitspraken vertrouwelijk worden behandeld.
    - Er worden geen uitspraken van de ene persoon gedeeld met de andere zodat het herleidbaar is tot de persoon die de uitspraak deed.
    - Resultaten zijn niet herleidbaar tot personen.
- De data van het onderzoek zorgvuldig worden behandeld en worden opgeslagen op locaties die aan wettelijke richtlijnen voldoen (AVG):
    - Er worden letterlijke transcripties van de interviews gemaakt. Er zit dus geen interpretatie van de onderzoeker in hetgeen dat wordt vastgelegd.
    - Na transcriptie worden de audio opnames vernietigd (zie ook UITNODIGING INTERVIEW).
    - Tijdens het onderzoek worden de transcripties opgeslagen op Surfdrive. Surfdrive voldoet aan Nederlandse privacy wetgeving[20] (Anon n.d.).

Geef de deelnemer voldoende informatie en vraag toestemming

De eerste toestemming die de deelnemers geven is als ze positief reageren op de uitnodiging via e-mail (UITNODIGING INTERVIEW). Tijdens de introductie wordt ingegaan op het onderzoek en kunnen extra vragen gesteld worden. Vervolgens wordt verzocht of het interview opgenomen mag worden. Dat is het tweede moment dat om toestemming wordt gevraagd.

Bescherm de privacy van de deelnemer

Sterk gerelateerd aan de vorige twee en heeft vooral te maken met het zonder toestemming naar de privézaken van een deelnemer kijken. Dat is in dit onderzoek niet aan de orde. Onderzoek vindt plaats op de case organisatie en er wordt alleen naar werkgerelateerde zaken gevraagd.

Misleid de deelnemer niet

Is niet van toepassing op dit onderzoek. Is bedoeld om te voorkomen dat een onderzoeker probeert meer medewerking te verkrijgen door het doel van het onderzoek anders uit te leggen als het is.

[20] SURF voldoet met SURFdrive aan de Nederlandse en Europese privacywetgeving (https://wiki.surfnet.nl/display/SURFdrive/Privacy+SURFdrive).

4. Resultaten

De deelvragen zoals benoemd onder METHODOLOGIE worden hieronder uitgewerkt. Eerst wordt de theorie uitgewerkt en wordt ingegaan op onderzoeksdata, metadata en datadocumentatie. Vervolgens wordt ingegaan op vindbaarheid, volgbaarheid, data opslag en archiveerbaarheid. Met de opgebouwde theorie wordt het WSS gevuld in de paragrafen erna.

4.1. Uitwerking theorie

Er zijn veertien interviews gehouden waarin de concepten uit FIGUUR 12 aan de orde kwamen. Deze concepten waren grotendeels de hoofdcategorieën van de codes waarin de interviews gecodeerd werden, bovendien kwamen in de interviews nog extra onderwerpen naar voren. Dit staat samengevat in TABEL 19.

Code categorieën met hun oorsprong en waar en hoe ze gebruikt zijn						
Code Categorie	Afkor-ting	Oor-sprong	# unieke quotes*	# Interviews	# subcodes	Bijlage
Onderzoeksdata	rd	Lit	110	14	10	ONDERZOEKSDATA
Metadata	md	Lit	82	14	9	Metadata
Datadocumentatie	dd	Lit	43	12	7	DATADOCUMENTATIE
Data opslag	opsl	Lit	83	14	7	DATA OPSLAG
Data delen	del	Lit	80	13	7	DATA DELEN
Data beveiliging**	bvlg	Lit	77	13	6**	DATA BEVEILIGING
Werkwijze	***	Int	53	12	8	WERKWIJZE
Vindbaarheid	vind	Lit	40	12	4	VINDBAARHEID
Volgbaarheid	vlg	Lit	62	13	6	VOLGBAARHEID
Archiveerbaarheid	archief	Lit	29	10	5	ARCHIVEERBAARHEID
Ontwerp	archit****	Int	30	10	3	ONTWERP

*In de tabel staat het aantal unieke quotes. Het kan zijn dat een quote vaker is gebruikt bij verschillende subcodes.
** Logging wordt bij beveiliging benoemd.
*** Verzameling van diverse subcodes, die staan benoemd in de bijlage.
**** Oorspronkelijk werd als code archit van architectuur gebruikt, later werd besloten dat ontwerp de lading beter dekte.
Folderstructuren en naamconventies komen terug bij vindbaarheid, volgbaarheid, metadata en werkwijze.
Versiebeheer staat onder volgbaarheid.
Doorzoekbare bron is niet expliciet uitgevraagd, zoeken en vinden is wel bevraagd onder vindbaarheid.

Tabel 19, Code categorieën met hun oorsprong en waar en hoe ze gebruikt zijn

De interviews zijn per concept uit FIGUUR 12 uitgewerkt in BIJLAGE UITWERKING INTERVIEWS. Deze bijlage kan ook als codeboek gebruikt worden. In de bijlage staan de definities en de werkwijze zoals de dagelijkse praktijk is bij de geïnterviewden uitgewerkt. Definities uit de literatuur voor onderzoeksdata, metadata, data opslag, vindbaarheid en volgbaarheid hoeven niet veranderd te worden, de eerste twee worden wel aangevuld en volgbaarheid wordt anders verwoord. Archiveerbaarheid wordt aangepast. Ook het concept van actieve data, krijgt een andere definitie en verdwijnt daardoor.

Dit hoofdstuk begint met de opgebouwde theorie uit de interviews en motiveert de definitieve formulering van de probleemstelling. Daarna wordt het WSS ingevuld op basis van de interviews en uiteindelijk naar de definitieve versie gebracht op basis van de expert review.

4.1.1. Onderzoeksdata, metadata en datadocumentatie

De definities van onderzoeksdata en metadata uit de literatuur worden bevestigd in de interviews. Bij onderzoeksdata wordt wel het belang benadrukt om de mogelijkheid te hebben om (delen van) het onderzoek te herhalen (ONDERZOEKSDATA DEFINITIE).

Een verdere belangrijke constatering is dat onderzoeksdata op drie niveaus bestaan (ONDERZOEKSDATA LEVELS):

1. Primary: De data zoals die direct na ontstaan zijn verkregen;
2. Intermediate: De data die bewerkt zijn en waar analyses op verricht worden;
3. Final: De data die gebruikt worden om de tabellen en figuren in de publicatie te vullen.

De eerdere definitie van metadata wordt uitgebreid met: Vaak bevatten de metadata gegevens die gebruikt kunnen worden bij het bepalen van de waarde van de verzamelde gegevens. Daarmee kunnen ze een juiste herhaling van het onderzoek bevorderen (METADATA DEFINITIE).

Het is belangrijk voor de herhaalbaarheid van het onderzoek dat, buiten de informatie die de metadata al bieden, beschreven is hoe de data op niveau 1 zijn verzameld en hoe ze zijn bewerkt om op niveau 2 en 3 te komen. Dit wordt inhoudelijk beschreven in datadocumentatie: Datadocumentatie tijdens onderzoek betekent het georganiseerd bijhouden van aantekeningen over hoe de data zijn verzameld, wat de resulterende databestanden zijn en hoe ze zijn verwerkt. (DATADOCUMENTATIE CONCLUSIE).

In de interviews wordt gemeld dat het meest volledige voorbeeld van datadocumentatie een elektronisch lablogboek (ELN) is. Hierin beschrijft men het verzamelen en de inhoud van onderzoeksdata, de metadata, de data bewerkingen en legt men soms ook links (een geschreven verwijzing of een hyperlink) naar databestanden (DATADOCUMENTATIE CONCLUSIE).

Alle data (onderzoeksdata, metadata en datadocumentatie) kunnen relevant zijn voor het herhalen van het onderzoek of van een deel van het onderzoek door dezelfde of door een andere onderzoeker. Welke data voor herhaalbaarheid precies relevant zijn, is ter beoordeling van de oorspronkelijke onderzoeker(s). Dit leidt tot de formulering relevante data in het lopende onderzoek i.p.v. actieve onderzoeksdata[21].

4.1.2. Volgbaarheid, vindbaarheid, data opslag en archiveerbaarheid

Volgbaarheid

Op basis van de interviews wordt de definitie van volgbaarheid aangepast: Volgbaarheid omvat de data provenance, het proces van het bijhouden van wijzigingen in de onderzoeksdata, inclusief de middelen waarin die wijzigingen worden bijgehouden (VOLGBAARHEID CONCLUSIE).

Volgbaarheid beschrijft hoe de overgangen tussen de verschillende niveaus van onderzoeksdata worden bewerkstelligd. RDM heeft mede als functie om (stappen in) het onderzoek te kunnen herhalen en hiervoor zijn metadata en datadocumentatie nodig, m.a.w. relevante data. Volgbaarheid is dan het bijhouden van de relevante data zodat (een deel van) het onderzoek herhaald kan worden (VOLGBAARHEID CONCLUSIE).

[21] Actieve data wordt door de geïnterviewden anders geïnterpreteerd: Data kunnen actief blijven doordat ze na het lopende onderzoek waarin ze ontstaan zijn, in een ander onderzoek opnieuw worden gebruikt (ONDERZOEKSDATA ACTIEF).

De invloed van versiebeheer, naamconventies en folderstructuren op de volgbaarheid van data is aanwezig, maar zwak. Een uitzondering is versiebeheer op code via daarvoor specifiek bedoelde systemen. Deze werkwijze vertoont ook een hoge mate van volgbaarheid. Goede beschrijvingen van onderzoeksdata en de vastlegging van hoe data zijn verzameld en achtereenvolgens zijn bewerkt in een ELN, zijn tezamen de sterkste vorm van volgbaarheid die is gevonden in dit onderzoek. Op deze manier werken, gebeurt maar bij twee van de veertien geïnterviewden. Eén andere geeft aan hier interesse in te hebben. (uit de BIJLAGE VOLGBAARHEID)

Vindbaarheid
Voor vindbaarheid blijft de definitie uit het literatuuronderzoek staan: De customers en participants van het werksysteem zijn in staat om de relevante data van het lopende onderzoek te vinden. Het verband tussen folderstructuren en vindbaarheid is sterker dan voor volgbaarheid. Door per onderzoek een bekende folderstructuur te gebruiken en die consequent te hanteren, kan men relevante data terugvinden. Een andere sterke verwantschap is het verband tussen een ELN en vindbaarheid: In een ELN kan men een verwijzing maken (bijvoorbeeld een snelkoppeling) naar een dataset. Dat maakt terugvinden eenvoudiger. Door de namen van de databestanden, die bij een onderzoeksstap horen, te voorzien van bepaalde beschrijvende parameters (denk aan een snelheid in het experiment, of een datum) kan daar op gezocht worden (vindbaarheid) maar het kan ook helpen met volgbaarheid (als er elke datum een andere snelheid wordt gebruikt bijvoorbeeld). (zie VINDBAARHEID CONCLUSIE).

De sterkste oplossing die zowel vindbaarheid als volgbaarheid goed dekt is het gebruik maken van een ELN gecombineerd met een consequent doorgevoerde folderstructuur per onderzoek.

Data opslag
Data worden opgeslagen op centrale oplossingen, share and sync oplossingen, externe datahouders en systemen specifiek bedoeld voor code (zie DATA OPSLAG CONCLUSIE).

Langdurige opslag en Archiveerbaarheid
Archiveerbaarheid wordt verschillend benaderd:

1. Het opslaan van relevante data uit het afgeronde onderzoek bij een daarvoor gespecialiseerde repository.
2. Het voor een relevante periode opslaan van de relevante data uit het lopende of afgeronde onderzoek op een plek waar deze data later teruggevonden kan worden, om (delen van) het onderzoek te kunnen herhalen of om de data voor ander onderzoek te gebruiken.

De eerste geldt als het lopende onderzoek is afgerond, de tweede is voor tijdens en na het lopende onderzoek. Voor dit onderzoek beperkt archiveerbaarheid zich tot de tweede eigenschap (zie ARCHIVEERBAARHEID CONCLUSIE).

In de tweede eigenschap gaat het om een locatie waar men ‘voor een relevante periode’ data kan opslaan (laten staan). Hoe lang een relevante periode is, kan alleen bepaald worden door de onderzoeker (in de case organisatie de PI).

Logischerwijs zou men in de datadocumentatie naar deze locaties verwijzen (datadocumentatie wordt immers tijdens het onderzoek gemaakt, er is dan nog geen sprake van archivering). Men slaat relevante data die langdurig bewaard moeten worden op (of laat ze staan) op externe datahouders. Voorwaarde hiervoor is dat er voldoende ruimte is, anders gaat men oude data opruimen (zie ook DATA OPSLAG en ONDERZOEKSDATA OPRUIMEN). Eén van de geïnterviewden is hierin specifiek en noemt een getal van standaard minimaal 8TB per onderzoek en meer op verzoek. Hij stelt ook een (hele) minimale eis aan de snelheid van het systeem, namelijk sneller dan een tapesysteem en noemt een

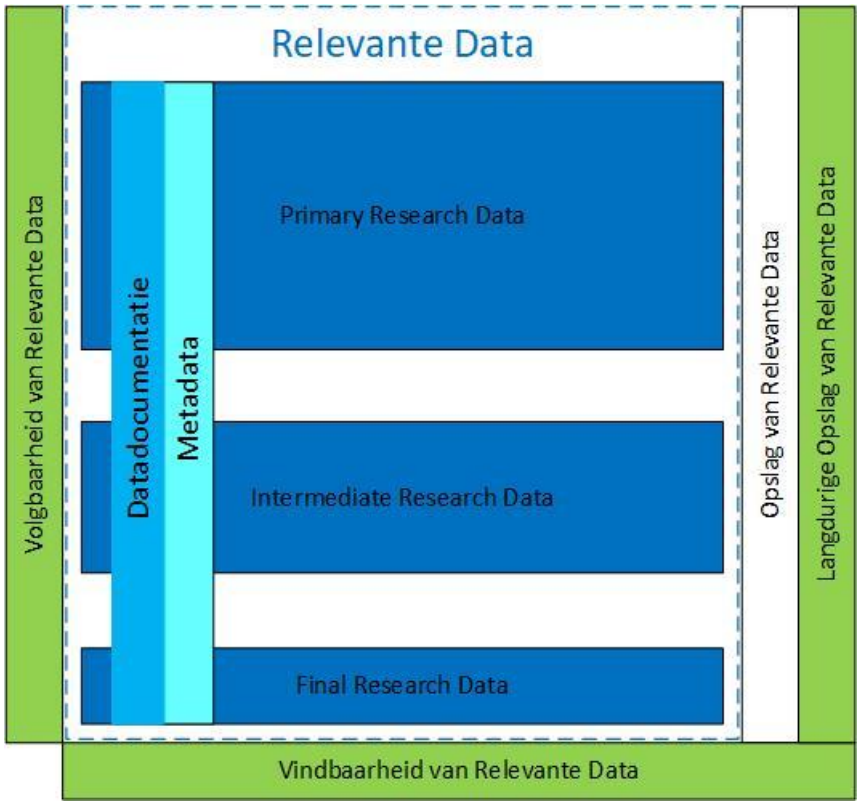
termijn van minimaal 10 jaar (ONTWERP CONCLUSIE). De termijn staat echter meestal niet vast en data worden opgeslagen zolang ze nu hebben. In Onderzoeksdata conclusie stelt men dat het bestaan van data pas eindigt als het wordt opgeruimd.

Analoog voor de formulering van relevante data wordt vanaf nu 'relevante periode' gebruikt. Dit resulteert in een laatste wijziging van de probleemstelling:

Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op de relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en voor relevante periode op te slaan zijn?

De opgebouwde theorie staat samenvattend afgebeeld in FIGUUR 13. De blauwe en groene componenten in de figuur vormen de hoofdcomponenten van de probleemstelling.

- In verschillende tinten blauw, tussen de onderbroken lijnen, staan de relevante data. De metadata en de datadocumentatie doorsnijden de verschillende niveaus van de onderzoeksdata, omdat ze deze ook op de verschillende niveaus beschrijven.
- De onderzoeksdata hebben verschillende formaten. Ruwe data is meestal het grootst en de data nodig voor figuren en tabellen in publicaties het kleinst. De schaal is illustratief en kan per onderzoek anders zijn.
- Rechts staat de opslag voor de data en daaraan vast (in het groen) de langdurige opslag. Uit het voorgaande blijkt dat dat vaak dezelfde datahouders zijn.
- Onder de relevante data en de opslag, staat vindbaarheid in het groen. De relevante data worden gevonden op de opslag.
- Links van de relevante data staat de volgbaarheid, vlak bij de datadocumentatie en metadata, die hier een belangrijke rol in spelen.



Figuur 13, grafische uitbeelding van de opgebouwde theorie na de empirische fase

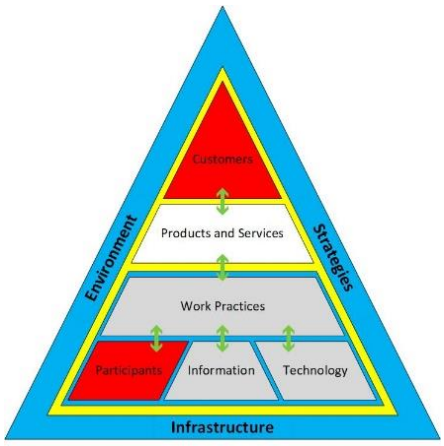
4.1.3. Het RDM werksysteem

Het RDM werksysteem wordt beschreven met het WSS. Het literatuuronderzoek zorgde voor een eerste versie van het WSS. Door de resultaten uit de interviews wordt in dit hoofdstuk een tweede versie gemaakt. Praktijk op de case organisatie en theorie uit de literatuur komen hierin samen. Daar waar theorie en praktijk elkaar bevestigen en/of aanvullen ontstaan ontwerprichtlijnen. Bovendien worden enkele ontwerprichtlijnen genoemd in de interviews. Het WSS wordt met dezelfde opzet als in de literatuurstudie gevuld.

De stakeholders van het RDM werksysteem

De participants voeren het werk uit (m.b.v. systemen). Ze zijn al eerder voor de case organisatie gespecificeerd in Selectie geïnterviewden:

- Principle Investigators (PI's);
- Onderzoekers;
- Data Stewards;
- IT ondersteuners (in de afdeling/sectie).



Figuur 14, stakeholders in het RDM Werksysteem

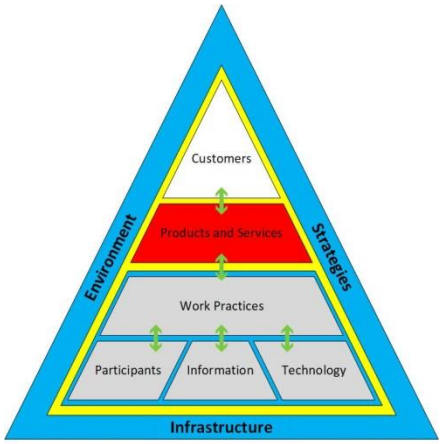
Kijkend naar de lijst met participants zijn dat de PI's en onderzoekers. (Zie ook PARTICIPANTS EN CUSTOMERS).

Products and services van het RDM werksysteem

Het proces van het verifiëren en/of aanvullen van de gevonden products and services uit de literatuur bij de case organisatie staat beschreven in PRODUCTS AND SERVICES. De resultaten staan in TABEL 20.

Belangrijkste aanpassingen t.a.v. de literatuurstudie zijn:

- Voor RDM wordt rekening gehouden met de relevante data voor het lopende onderzoek i.p.v. actieve data;
- In plaats van archivering, gaat het om langdurige opslag die relatief snel te benaderen is.



Figuur 15, products and services in het RDM Werksysteem

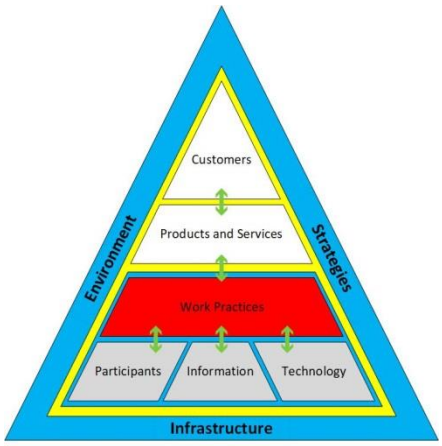
Products and services, uiteindelijke lijst					
#	**Products and services**	**VDB**	**VLG**	**LO**	
1.	Het werksysteem verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	X			Bekende locatie maakt vinden makkelijker.
2.	Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen externe datahouders en een centrale opslag.	X			Data staan op de centrale opslag na synchronisatie. Dat verhoogt vdb.
3.	Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek aan te houden waarbij formats in templates worden beschreven.	X	X		Relevante data en folderstructuren en naamconventies tezamen vormen een goede basis voor vdb en vgb.
4.	Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.	X	X		Vdb en vgb zijn onderdeel van de stelling.
5.	Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.
6.	Het werksysteem biedt de mogelijkheid om geautomatiseerd metadata aan te maken, bij voorkeur m.b.v. een template.	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
7.	Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).	X	X		Om data te kunnen vinden en volgen is een interface om mee te zoeken noodzakelijk.
8.	Het werksysteem maakt het mogelijk data te vernietigen en logt deze bewerking.	X	X		De laatste bewerking op data. Door de logging weet je dat de data verdwenen zijn en niet gevonden kunnen worden
9.	Het werksysteem biedt de mogelijkheid een ELN te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.	X	X		ELN kan voor vdb en vlg zorgen.
10.	Het ELN in het werksysteem biedt de mogelijkheid om o.a. activiteiten en keuzes te loggen	X	X		Helpt vdb en vgb als je bijv. logt waar je iets hebt opgeslagen of wat je met een dataset hebt gedaan.
11.	Het ELN in het werksysteem biedt de mogelijkheid om te linken naar relevante data in het lopende onderzoek.	X	X		Vgb want linken naar data die transformaties beschrijven.
12.	Het ELN in het werksysteem kan specifiek gemaakt worden voor bepaalde vakgebieden		X		Bijvoorbeeld een procesbeschrijving in chemische wetenschappen.
13.	Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek te beveiligen.		X		Het verschaffen van autorisaties, i.c.m. logging bevordert de volgbaarheid.
14.	Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan. Databewerkingen worden gelogd.		X		Op voorwaarde van geschikte logging volgbaar.
15.	Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.		X		Helpt datatransformaties inzichtelijk te maken.

#	Products and services	VDB	VLG	LO	
16.	Het werksysteem biedt een opslagplaats voor koude relevante data die voldoende snel (sneller dan tape) opgehaald kan worden			X	Langdurige opslag van data in lopend onderzoek en daarna.
VDB: Geeft antwoord op de vraag: Waar zijn de relevante data uit het lopende onderzoek (fysiek en/of logisch)? VLG: Geeft antwoord op de vraag: Kan ik de wijzigingen die de relevante onderzoeksdata ondergaan volgen op basis van de beschikbare informatie? LO: Geeft antwoord op de vraag Kan ik relevante data uit het lopende onderzoek ergens langdurig opslaan?					

Tabel 20, Products and services, uiteindelijke lijst.

Work practices van het werksysteem

Work practices werden eerder afgeleid van de in de literatuur gevonden products and services. Ze kunnen nu bepaald worden op basis van de uitkomsten van de interviews (en zijn dan per definitie specifiek voor de case organisatie). Vervolgens kan worden gecontroleerd of de bevindingen uit de literatuur zijn afgedekt. Als ze niet afgedekt zijn, dan worden ze toegevoegd. Het resultaat staat in TABEL 21. De uitwerking staat beschreven in ACTIVITEITEN.



Figuur 16, activiteiten in het RDM Werksysteem

Work practices, uiteindelijke lijst						
#	**Work practices**	**VDB**	**VLG**	**LO**		
1.	Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan	X				LIT
2.	Het werksysteem keurt de aanvraag en kent ruimte toe of niet	X				LIT
3.	Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte	X				LIT
4.	De researcher verplaatst (meta)data op de geboden opslagruimte	X				LIT
5.	De onderzoeker ruimt data op	X			Als je weet dat het is opgeruimd vdb	RD
6.	De onderzoeker verwijst naar een resulterend databestand met een link in de datadocumentatie	X			Beschrijft vdb	DD
7.	De onderzoeker slaat voldoende relevante data uit het lopende onderzoek op op een plek waar hij/zij het kan terugvinden	X			Vdb het staat er letterlijk	ARC
8.	De onderzoeker bewerkt onderzoeksdata op 3 verschillende niveaus	X	X		Om te kunnen bewerken, moet data vindbaar zijn. Voor verschillende niveaus volgbaar	RD
9.	De onderzoeker maakt metadata aan/gebruikt metadata	X	X		Metadata kunnen helpen bij vdb en vgb	MD
10.	De onderzoeker genereert geautomatiseerd metadata	X	x		Metadata kunnen helpen bij vdb en vgb	MD

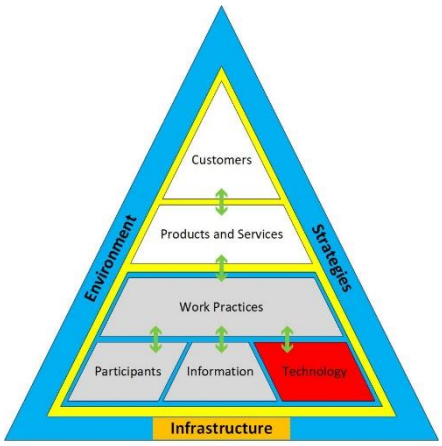
#	Work practices	VDB	VLG	LO		
11.	De onderzoeker beschrijft metadata in lablogboeken	X	X		Metadata kunnen helpen bij vdb en vgb	MD
12.	De PI adviseert een folderstructuur of legt deze op	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
13.	De PI adviseert een naamconventie of legt deze op	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
14.	De onderzoeker gebruikt een folderstructuur	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
15.	De onderzoeker gebruikt een naamconventie	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
16.	De onderzoeker houdt georganiseerd aantekeningen bij over de resulterende databestanden (wat ze zijn, waar ze staan e.d.)	X	X		Vdb staat in de stelling. weten hoe elk databestand is bewerkt is vgb	DD
17.	Er worden backups gemaakt van relevante data voor het lopende onderzoek	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.	DO
18.	De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verzameld		X		Hoe het wordt verzameld, kan iets zeggen over vgb	DD
19.	De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verwerkt		X		weten hoe elk databestand is bewerkt is vgb	DD
20.	De relevante data bewerkingen en benaderingen worden gelogd		X		Zo kun je volgen wat er met de data gebeurt	DB
21.	De onderzoeker pas versiebeheer toe		X		Versiebeheer helpt bij vgb	VLG
22.	De onderzoeker herhaalt (delen van) het onderzoek	X	X	X	Hiervoor is vdb, vgb en evt. lo nodig	RD

LIT, THEORETISCH KADER
RD, ONDERZOEKSDATA
DD, DATADOCUMENTATIE
ARC, ARCHIVEERBAARHEID
MD, METADATA
DO, DATA OPSLAG
DB, DATA BEVEILIGING
VLG, VOLGBAARHEID

Tabel 21, Work practices, uiteindelijke lijst

Technology en infrastructure van het RDM werksysteem

Technology bestaat met name uit hulpmiddelen die het werk ondersteunen. In de interviews worden voor technology de volgende hulpmiddelen genoemd:



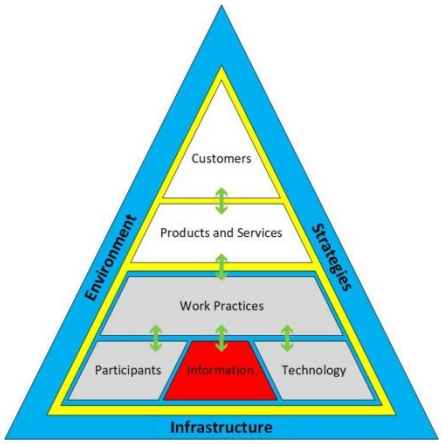
Figuur 17, technology in het RDM Werksysteem

- De diverse toegepaste opslagsystemen (zoals benoemd in DATA OPSLAG CONCLUSIE);
- Data delen vanaf de verschillende toegepaste opslagsystemen (zoals benoemd in DATA DELEN CONCLUSIE);
- Data beveiligingstechnieken (zoals benoemd in CONCLUSIE DATA BEVEILIGING);
- Tooling om versiebeheer van code geautomatiseerd uit te voeren;
- ELN's (en ook papieren logboeken).

Technology als versiebeheer staat naast infrastructure als opslagsystemen. De praktijk bij de case organisatie spreekt de samenvoeging technology/infrastructure niet tegen.

Information binnen het werksysteem

De volgende opmerkingen t.a.v. information worden gemaakt in de interviews (INFORMATION):

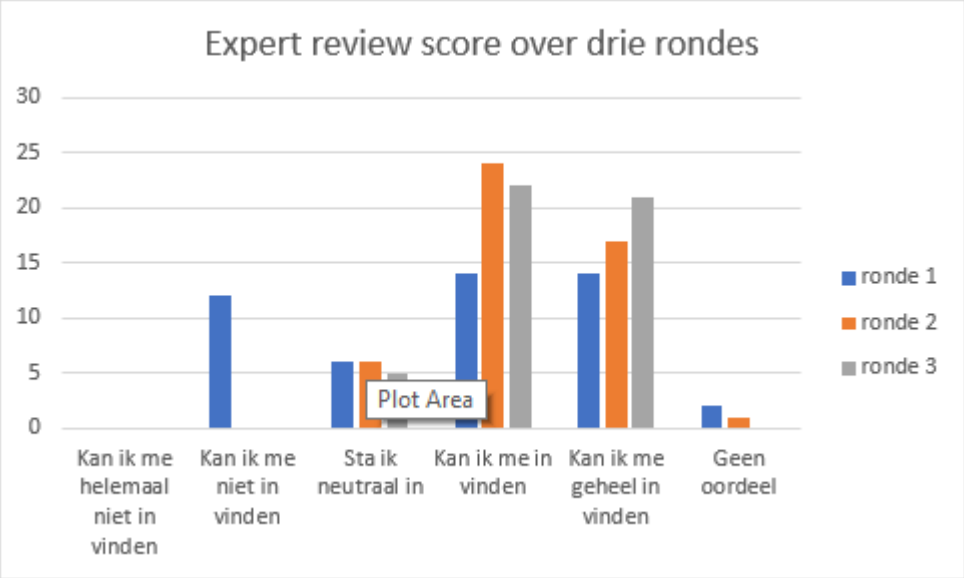


Figuur 18, informatie in het RDM Werksysteem

- Adviezen van data stewards aan onderzoekers (inclusief het opmaken van een data management plan);
- Het data management plan, met de planning van het databeheer.
- Adviezen van senior researchers (meestal de PI) aan de onderzoekers (o.a. voor folderstructuren en naamconventies);
- Versiebeheer handmatig uitgevoerd;
- Ruimte aanvragen door een PI;
- Overzichten uit logging.

4.1.4. Expert review

De products and services die na de interviews waren beschreven (TABEL 20) werden aan drie experts voorgelegd. De expert review bereikte na de tweede ronde reeds consensus. In de tweede ronde werden verfijningsvoorstellen gedaan, die in de derde voor het grootste deel werden bevestigd. De review was daarmee tot een einde gekomen. Het verschuiven van de meningen en het toegroeien naar consensus wordt beschreven in FIGUUR 19. Het hele proces staat beschreven in BIJLAGE EXPERT REVIEW. De resultaten van de expert review staan in TABEL 22.



Figuur 19, verschuiven van meningen en toegroeien naar consensus in expert review over drie rondes

De resultaten staan in TABEL 22.

Products and services, uiteindelijke lijst na expert review	
#	**Products and services**
1.	Het werksysteem verzekert dat er in het lopende onderzoek voldoende, door de onderzoeker te bepalen, betrouwbare opslagruimte per onderzoek beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.
2.	Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen onderzoeksdatahouders en een centrale opslag.
3.	Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie, waarbij versies worden bijgehouden van de verschillende replica's en waarbij een bepaalde versie van een replica teruggesynchroniseerd kan worden.
4.	Het werksysteem biedt een opslagplaats waar voor een relevante periode de relevante data van een onderzoek voldoende snel opgehaald kunnen worden. Voldoende snel is afhankelijk van hoeveelheden relevante data en de mate waarin ze nog gebruikt worden.
5.	Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek te ondersteunen waarbij formats in templates beschreven kunnen worden.
6.	Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op relevante data te kunnen toepassen.
7.	Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de relevante data zo op te slaan, dat duidelijk is dat ze bij elkaar horen.
8.	Het werksysteem maakt het mogelijk relevante data en alle afgeleiden hiervan, te vernietigen en logt deze activiteit.
9.	Het werksysteem levert voorzieningen om metadata te genereren met zo min mogelijk inspanning van de onderzoeker, bij voorkeur geautomatiseerd op basis van een template of algoritmen, maar ook met voldoende ruimte voor eigen invulling.
10.	Het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via meerdere interfaces (zoals mens-machine, machine-machine).
11.	Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.
12.	Het werksysteem bewaakt de integriteit van de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden in logging.
13.	Het werksysteem kan integreren met bestaande, veelgebruikte, ELN toepassingen.
14.	Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen.
15.	Het werksysteem biedt een ELN dat de mogelijkheid heeft om een link (verwijzing, 'shortcut') vast te leggen naar relevante data in het lopende onderzoek.
16.	Het werksysteem biedt een ELN aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden, zoals het tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.

Tabel 22, products and services, uiteindelijke lijst na expert review

4.1.5. Ingevuld WSS van het RDM werksysteem

In TABEL 23 staat de uiteindelijke versie van het WSS getoond. Ten opzichte van het eerdere WSS, is het wat specifieker geworden. Dat geldt met name voor de work practices, information, technology en products and services. Een belangrijke aanname aan het begin van het onderzoek, was dat infrastructure en technology gecombineerd konden worden. In het geval van een RDM werksysteem, is ICT infrastructuur een belangrijk component dat RDM mogelijk maakt. De technology-onderdelen in TABEL 23, hebben inderdaad een sterke relatie met ICT infrastructuur en de keuze lijkt daarom valide.

Een tweede aspect dat opvalt is dat het uiteindelijke WSS een combinatie is van bestaande technieken en werkwijzen en gewenste technieken en werkwijzen. Dat betekent dat het verschil tussen wat er is en wat er wordt gewenst, niet in één oogopslag zichtbaar is. Als men een geheel nieuw werksysteem wil inrichten, is dat geen probleem. Als men wil voortborduren op bestaande werkwijzen en technieken, dan is het wellicht verstandig expliciet te maken wat er is en wat extra gewenst is.

Voor een werksysteem vormen environment, strategies en infrastructure onderdelen van de buitenwereld. In het WSS worden ze niet meegenomen. Voor infrastructure geldt dat minder door de koppeling aan technology. Veel culturele aspecten (zoals benoemd in WERKWIJZE) zijn met name impliciet in de products and services en work practices verwerkt. In een omgeving waar werk sterk gediversifieerd is en mensen met elkaar en met systemen moeten werken om tot een resultaat te komen, lijkt een explicietere nadruk op cultuur en werkwijzen van belang. Bijvoorbeeld technologie die niet eenvoudig is aan te leren en de onderzoeker onvoldoende werk uit handen neemt, zal niet geadopteerd worden.

<table>
<tr><th colspan="3">Work system snapshot voor managing active data die vindbaar, volgbaar en archiveerbaar zijn</th></tr>
<tr><th>Customers</th><th colspan="2">Products & Services</th></tr>
<tr><td>• Principle Investigators (PI’s);
• Onderzoekers;
• Computational stakeholders.</td><td colspan="2">• Het werksysteem verzekert dat er in het lopende onderzoek voldoende, door de onderzoeker te bepalen, betrouwbare opslagruimte per onderzoek beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.
• Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen onderzoeksdatahouders en een centrale opslag.
• Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie, waarbij versies worden bijgehouden van de verschillende replica’s en waarbij een bepaalde versie van een replica teruggesynchroniseerd kan worden.
• Het werksysteem biedt een opslagplaats waar voor een relevante periode de relevante data van een onderzoek voldoende snel opgehaald kunnen worden. Voldoende snel is afhankelijk van hoeveelheden relevante data en de mate waarin ze nog gebruikt worden.
• Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek te ondersteunen waarbij formats in templates beschreven kunnen worden.
• Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op relevante data te kunnen toepassen.
• Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de relevante data zo op te slaan, dat duidelijk is dat ze bij elkaar horen.
• Het werksysteem maakt het mogelijk relevante data en alle afgeleiden hiervan, te vernietigen en logt deze activiteit.
• Het werksysteem levert voorzieningen om metadata te genereren met zo min mogelijk inspanning van de onderzoeker, bij voorkeur geautomatiseerd op basis van een template of algoritmen, maar ook met voldoende ruimte voor eigen invulling.
• Het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via meerdere interfaces (zoals mens-machine, machine-machine).
• Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.
• Het werksysteem bewaakt de integriteit van de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden in logging.
• Het werksysteem kan integreren met bestaande, veelgebruikte, ELN toepassingen.
• Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen.
• Het werksysteem biedt een ELN dat de mogelijkheid heeft om een link (verwijzing, ‘shortcut’) vast te leggen naar relevante data in het lopende onderzoek.
• Het werksysteem biedt een ELN aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden, zoals het tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.</td></tr>
<tr><th colspan="3">Work practices</th></tr>
<tr><td colspan="3">• Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan
• Het werksysteem keurt de aanvraag en kent ruimte toe of niet;
• Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte.
• De researcher verplaatst (meta)data op de geboden opslagruimte;
• De onderzoeker ruimt data op
• De onderzoeker verwijst naar een resulterend databestand met een link in de datadocumentatie
• De onderzoeker slaat voldoende relevante data uit het lopende onderzoek op op een plek waar hij/zij het kan terugvinden
• De onderzoeker bewerkt onderzoeksdata op 3 verschillende niveaus
• De onderzoeker maakt metadata aan/gebruikt metadata
• De onderzoeker genereert geautomatiseerd metadata
• De onderzoeker beschrijft metadata in lablogboeken
• De PI adviseert een folderstructuur of legt deze op
• De PI adviseert een naamconventie of legt deze op
• De onderzoeker gebruikt een folderstructuur
• De onderzoeker gebruikt een naamconventie
• De onderzoeker houdt georganiseerd aantekeningen bij over de resulterende databestanden (wat ze zijn, waar ze staan e.d.)
• Er worden backups gemaakt van relevante data voor het lopende onderzoek
• De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verzameld
• De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verwerkt
• De relevante data bewerkingen en benaderingen worden gelogd
• De onderzoeker pas versiebeheer toe
• De onderzoeker herhaalt (delen van) het onderzoek</td></tr>
<tr><th>participants</th><th>Information</th><th>Technologies</th></tr>
<tr><td>• Principle Investigators (PI’s);
• Onderzoekers;
• Data Stewards;
• IT ondersteuners (in de afdeling/sectie).
• Computational stakeholders.</td><td>• Adviezen van data stewards aan onderzoekers (inclusief het opmaken van een data management plan);
• Het data management plan, met de planning van het databeheer.
• Adviezen van senior researchers (meestal de PI) aan de onderzoekers (o.a. voor folderstructuren en naamconventies);
• Versiebeheer handmatig uitgevoerd;
• Ruimte aanvragen door een PI;
• Overzichten uit logging.</td><td>• De diverse toegepaste opslagsystemen (zoals benoemd in Data opslag conclusie);
• Data delen vanaf de verschillende toegepaste opslagsystemen (zoals benoemd in Data delen conclusie);
• Data beveiligingstechnieken (zoals benoemd in Conclusie data beveiliging);
• Versiebeheer geautomatiseerd uitgevoerd;
• ELN’s (en ook papieren logboeken).</td></tr>
</table>

Tabel 23, ingevuld WSS, uiteindelijke versie

Expres leeggelaten i.v.m. formaat WSS op vorige pagina.

5. Discussie, conclusies en aanbevelingen

In dit onderzoek wordt uiteindelijk antwoord gegeven op de vraag:
Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en voor relevante periode op te slaan zijn?
Deze probleemstelling is het gevolg van een ontwikkeling tijdens het onderzoek. De formuleringen 'relevante data voor lopende onderzoeken' en 'voor relevante periode opslaan' zijn preciseringen die tijdens de empirische fase ontstonden en beter passen bij de werkelijkheid van de case organisatie (zie ook BIJLAGE ONTWIKKELING VAN DE PROBLEEMSTELLING).

5.1. Discussie aanpak

Er is gekozen om DSR te gebruiken om tot een ontwerp (construct) te komen in de vorm van een WSS wat een onderdeel is van de WSM. Vinden, volgen en archiveren van relevante data is een onderdeel van RDM. Om de eerste versie van het WSS in te vullen is er theorie opgebouwd op basis van literatuuronderzoek. Dit leidde tot veertien onderwerpen van gesprek in de interviews (zie ook FIGUUR 12). Voor de tweede versie zijn semigestructureerde interviews uitgevoerd onder een populatie van onderzoekers, data stewards en een ondersteuner. De uitkomsten van de products and services zijn gecontroleerd en becommentarieerd door experts m.b.v. de Delphi methode. Dit leverde een derde en uiteindelijke versie van het WSS op. Het aantal interviews en het aantal besproken concepten, leidden tot een grote hoeveelheid informatie. Er is het nodige aan methoden en technieken gecombineerd om tot een eindversie te komen. Er wordt kort op de uitvoering van elke methode of techniek ingegaan.

5.1.1. DSR, Semigestructureerde interviews en grounded theory

Volgens Gregor en Hevner vindt DSR plaats in multidisciplinaire teams. Zij zeggen hierover: :"[] the teams' cognitive skills (e.g., creativity and reasoning) in designing innovative solutions; and the teams' social skills in bringing together all of the individual members' collective intelligence via effective teamwork." (Gregor and Hevner 2013:p7). Creativiteit en innovatieve oplossingen, suggereren een hoge mate van vrijheid in de aan te dragen oplossingen. 'Collective intelligence en 'effective teamwork' suggereren veel interactie. Op basis van TABEL 18 kan geconcludeerd worden dat het ontwerpteam multidisciplinair was (met onderzoekers uit tien verschillende disciplines en data stewards uit twee verschillende faculteiten). De stakeholders zijn, met name om planningsredenen, los van elkaar geïnterviewd en hebben niet kunnen reageren op elkaars uitspraken. In de praktijk bleek dat semigestructureerde interviews, waarbij alle belangrijke gevonden concepten uit de literatuur aan bod kwamen en er ook de mogelijkheid was voor veel eigen inbreng (zoals onder WSS beschreven is dat belangrijk) een goede aanpak was om relevante informatie te verzamelen. De theorie werd opgebouwd door interviews te structureren en verwantschappen te zoeken (grounded theory). Hiermee konden de bevindingen uit de literatuur geverifieerd worden, konden deze bevindingen waar nodig aangepast worden en konden er nieuwe bevindingen toegevoegd worden die het WSS zouden verrijken. De keuze voor DSR voor dit vraagstuk is terecht en in de omstandigheden van de case organisatie was de keuze voor semigestructureerde interviews in de praktijk het best.

5.1. Expert review met Delphi methode

Een kleine groep experts heeft de resultaten (products and services) beoordeeld op ICT infrastructuur, processen en innovatie (de eerste twee met name aansluitend op WSS en de laatste op DSR). Zij zagen elkaars (schriftelijke en anonieme) commentaar en konden daar op reageren. Dit gaf wat extra interactie. Dit blijkt een goede manier te zijn om op basis een bestaande set aan bevindingen een 'discussie' aan te gaan. Iedereen krijgt daarin evenveel ruimte en men kan vanuit eigen ervaring reageren op elkaars commentaar. Dit werkt waarschijnlijk goed als de groep niet te groot is en de ervaringen niet te ver uiteenlopen. Consensus was in dit geval snel bereikt. Geen van de resultaten werd verworpen. Aanpassingen betroffen met name verduidelijkingen en verscherping van de formuleringen.

5.1.1. WSS

Het WSS representeert een moment in tijd en beschrijft hoe mensen, informatie en systemen samenwerken om tot een resultaat te komen. Het is een model en laat daarom altijd onderdelen van de werkelijkheid weg. Wat bijvoorbeeld veel terug kwam in de interviews was het individualisme van de onderzoekers en hun pragmatisme. Men richt onderzoek op eigen wijze in, documenteert het op eigen wijze en gebruikt alleen ondersteunende middelen die eenvoudig zijn in gebruik en zo min mogelijk afleiden van het wetenschappelijke probleem. Dit zou in de representatie van een werksysteem in de omgeving terecht komen, maar die omgeving is geen onderdeel van het WSS. Toch is dit een belangrijk aspect voor een ontwerp. Het werksysteem moet bruikbaar zijn voor een zeer diverse groep en het moet eenvoudig toepasbaar zijn (anders kiest men iets anders). Deze eigenschap wordt nu impliciet verwerkt in sommige products and services, bijvoorbeeld metadata genereren met zo min mogelijk inspanning van de onderzoeker (TABEL 23). Dit voorbeeld toont aan dat het WSS wellicht niet in alle situaties de juiste nuances legt. Het resultaat toont met name een globale werkwijze en enkele kaders en eigenschappen waaraan het werksysteem zou moeten voldoen. Dat is mogelijk ook het gevolg van een grote scope van veertien besproken concepten. Het WSS heeft in die zin meer de eigenschap van een globaal ontwerp of een architectuurplaat dan van een detail ontwerp.

5.1.2. Concluderend bij discussie aanpak

Invulling van DSR met zowel de literatuurstudie, als de keuze voor theorieopbouw met semigestructureerde interviews via grounded theory i.p.v. focusgroepen (met meer interactie) leidt tot een globaal afkaderend WSS. Het WSS als moment in tijd van een werksysteem, lijkt sowieso meer geschikt voor een globale beschrijving dan een detail ontwerp. De bevindingen zijn het gevolg van met name het zoeken naar overlap in literatuur en in uitspraken van verschillende stakeholders. De kans dat de bevindingen inhoudelijk kloppen wordt met deze aanpak verhoogd en de interne en externe validiteit is op deze manier voor dit soort onderzoek hoog.[22]

5.2. Discussie inhoudelijk

Literatuur over RDM richt zich met name op de dataopslag na het onderzoek. Literatuur over dataopslag tijdens lopend onderzoek is beperkt voor handen. Het vindbaar en volgbaar houden van data in lopend onderzoek is niet eerder op deze wijze specifiek onderzocht. Ook langdurige opslag in tegenstelling tot archivering is, voor zover bekend, specifiek voor dit onderzoek. Bevindingen uit de literatuur zijn afgeleid van uitspraken in literatuurbronnen, die van toepassing zijn op vindbaarheid,

[22] Dit wil overigens niet zeggen dat de interne en externe validiteit van dit onderzoek hoog zijn. Alleen dat getracht is ze te maximaliseren.

volgbaarheid en archivering. Deze bevindingen werden in de empirische fase minimaal bevestigd en soms ook uitgebreid. De eigenschappen uit de literatuur bleken generaliseerbaar en van toepassing op de onderzochte instelling.

5.3. Conclusies

Op een globaal niveau geeft het onderzoek een goed antwoord op de onderzoeksvraag. De input voor de theorie en het daaruit voortvloeiende WSS, is via drie verschillende onderzoeksmethoden tot stand gekomen. Een gedegen literatuuronderzoek gevolgd door het op verschillende manieren uitvragen van mensen met verschillende functies. Het WSS is daardoor waarschijnlijk een goed uitgangspunt voor detailontwerpen. De belangrijkste onderdelen van de theorie die in dit onderzoek is opgebouwd zijn:

- Onderzoekers hebben een hoge mate van vrijheid bij de invulling van RDM voor hun onderzoek. Zij kunnen dit zelf bedenken of krijgen hier een advies over waar ze van af kunnen wijken. Het komt zelden voor dat de invulling (volledig) wordt opgelegd. Het is dus aan de onderzoeker wat moet worden opgeslagen, waar dit moet worden opgeslagen en voor hoe lang. Hij/zij bepaalt wat relevante data zijn en wat een relevante periode is om deze data op te slaan;
- Een belangrijke functie van RDM is om relevante data op te slaan zodat de mogelijkheid om (delen van) het onderzoek te herhalen aanwezig is;
- Relevante data zijn de onderzoeksdata, de metadata en de datadocumentatie die tezamen benodigd zijn om (delen van) het onderzoek te herhalen;
- Onderzoeksdata hebben verschillende niveaus (die verlopen van direct na verzamelen tot de data die nodig is voor de tabellen en figuren in de publicatie);
- Volgbaarheid beschrijft overgangen binnen en tussen de niveaus en waar dat geregistreerd wordt (in metadata en datadocumentatie);
- De volgbaarheid van de relevante data wordt voornamelijk bevorderd door gebruik te maken van een ELN. Voor codedata werkt versiebeheer goed door gebruik te maken van systemen die specifiek bedoeld zijn voor codebeheer (opslag, versiebeheer en delen);
- Met name folderstructuren en ELN's zorgen voor vindbaarheid;
- De sterkste oplossing die zowel vindbaarheid als volgbaarheid goed dekt is het gebruik maken van een ELN gecombineerd met een consequent doorgevoerde folderstructuur per onderzoek.
- Voor lopend onderzoek is niet zozeer archivering, maar opslag van belang. Opslag die een relevante periode beschikbaar is (gedurende het onderzoek en voor een periode die als nuttig wordt beoordeeld door de hoofdonderzoeker erna);
- Deze opslag kan decentrale datahouders betreffen en/of centraal door de instelling aangeboden opslag. Het is een wens om deze data te kunnen synchroniseren tussen decentrale datahouder en centrale opslag;
- Voor functionaliteit die het wetenschappelijke werk ondersteunt, lijken onderzoekers een voorkeur te hebben voor tooling die dit laagdrempelig en geautomatiseerd voor ze regelt. Een succesvol voorbeeld zijn de opslag systemen voor code die geautomatiseerd versiebeheer toepassen.

Deze aspecten komen vooral terug in de products and services zoals ze beschreven zijn in het WSS in TABEL 23.

5.4. Aanbevelingen voor de praktijk

Een logisch gevolg van een globaal ontwerp, zijn detailontwerpen. Als men bijvoorbeeld besluit om de projectopslag sync and share functionaliteit te geven, kan men het WSS als beginpunt gebruiken.

Het WSS kan ook als kader gebruikt worden. Bij bijvoorbeeld de aanschaf van een dienst die te maken heeft met beheer van wetenschappelijke data, kan men controleren of het past in de onderzoeksgewoonten en -werkwijzes van de organisatie.

Wetenschappers moeten voor aanvang van onderzoek een datamanagement plan maken om subsidie te krijgen voor hun onderzoek. Het WSS geeft extra kaders om de relevante data vindbaar en volgbaar te maken en voor (langdurige) opslag. Data stewards zouden dit samen met de wetenschappers in enkele praktijkcases kunnen uitproberen. Dit geeft ook weer extra input voor onderdelen van het WSS, zodat de inhoud verder verrijkt kan worden.

Er wordt door enkele wetenschappers aangegeven dat ze graag centrale opslag zouden willen benaderen met share and sync functionaliteit. Een combinatie van projectdrive en Surfdrive waarbij men data van decentrale datahouders kan synchroniseren met centrale opslag. Dit zou bijvoorbeeld op kleine schaal uitgeprobeerd kunnen worden. (Een dergelijke benadering kan voor diverse products and services toegepast worden.)

5.5. Aanbevelingen voor verder onderzoek

De bevindingen in de literatuur, gebaseerd op meerdere academische instellingen, bleken vrij goed generaliseerbaar naar de case organisatie. Het is te verwachten dat de bevindingen van de case organisatie t.a.v. vindbaarheid, volgbaarheid en langdurige opslag ook weer generaliseerbaar zijn naar andere universiteiten. Dit zou onderzocht kunnen worden. De products and services zouden met kwantitatief onderzoek (bijvoorbeeld een survey) geverifieerd kunnen worden bij een grotere populatie van onderzoekers, data stewards en (ICT) ondersteuners. Als dit ook bij andere universiteiten wordt gedaan, zou dit het eindresultaat extern generaliseerbaar maken.

Via case studies zouden verschillende products and services uitgewerkt kunnen worden tot een ontwerp dat in de praktijk wordt uitgeprobeerd bij een afdeling. Men zou hier DSR voor kunnen gebruiken en de bevindingen van dit onderzoek als basis kunnen nemen.

Er zijn diverse kaders en richtlijnen voor RDM. Voor privacy bijvoorbeeld de AVG wetgeving en voor onderzoeksubsidie het DMP. Het WSS zou hierop aangescherpt kunnen worden na nader onderzoek.

6. Procesreflectie

In deze reflectie wordt kort ingegaan op begeleiding, planning, verloop en leerpunten.

Begeleiding

De begeleiding was intern. Dat heeft veel voordelen en die zitten voornamelijk in de eenvoudige afstemming en de bekendheid met de materie. Mogelijk heeft de onbekendheid met het afstudeerproces bij de OU er wel toe geleid dat ik veel vrijheid nam om te grootse plannen te maken.

Planning

Het scopen van de onderzoeksvraag en de aanpak pasten naar mijn mening prima in een vijftien punten opdracht. Literatuuronderzoek was meer werk omdat er extra punten gehaald moesten worden. Interviews bleken meer werk dan verwacht. Expert review ging volgens plan.

Verloop

Het verwerken van interviews onderschatte ik: voorbereiden, uitvoeren, transcriberen en coderen. Vervolgens meer interviews uitvoeren, bestaande codes aanpassen en de waarde van de relaties tussen de codes onderbouwen. Uiteindelijk wordt dan theorie opgebouwd en op degelijke wijze beschreven. Tezamen levert het bijna 80 pagina's aan bijlages op en vijf theoriepagina's in het rapport.

Leerpunten

Uiteindelijk heeft het me veel gebracht. Niet alleen een keer zelf onderzoek doen en de academische cyclus doorlopen, maar ook de gebruikte methodes. In mijn werk kijk ik met een werksysteembril naar werkzaamheden, probeer ik met collega's vanuit verschillende disciplines naar een probleem te kijken en heb ik in verschillende trajecten inmiddels de documentatie gecodeerd en in een caqdas pakket bijgehouden.

Beschouwend

Veel is goed gegaan en wat mis ging in de planning had vooral met mijn onervarenheid te maken. Sowieso is het een bijzondere periode geweest met een pandemie, thuisonderwijs en thuiswerk.

Referenties

Akers, Katherine G., and Jennifer Doty. 2013. “Disciplinary Differences in Faculty Research Data Management Practices and Perspectives.” *International Journal of Digital Curation* 8(2):5–26.

Alexogiannopoulos, E., S. McKenney, and M. Pickton. 2010. *Research Data Management Project: A DAF Investigation of Research Data Management Practices at The University of Northampton*.

Alter, Steven. 2006. *The Work System Method: Connecting People, Processes, and IT for Business Results*. Work System Method.

Alter, Steven. 2008. “Defining Information Systems as Work Systems: Implications for the IS Field.” *European Journal of Information Systems* 17(5):448–69.

Alter, Steven. n.d. “Steven ALTER | Professor of Information Systems | University of San Francisco, CA | USFCA | School of Management.” Retrieved May 15, 2021 (https://www.researchgate.net/profile/Steven-Alter).

Anon. 2015. “Co-Occurrence And Correlation: Challenges Using ATLAS.Ti.”

Anon. 2020. “Data Provenance - Research Data Management - LibGuides at VU Amsterdam.” Retrieved September 18, 2020 (https://libguides.vu.nl/rdm/data-provenance).

Anon. n.d. “Data Documentation - WUR.” Retrieved September 16, 2020a (https://www.wur.nl/en/Value-Creation-Cooperation/WDCC/Data-Management-WDCC/Doing/Data-Documentation.htm).

Anon. n.d. “Veelgestelde Vragen over SURFdrive | SURF.Nl.” Retrieved August 7, 2020c (https://www.surf.nl/bewaar-en-deel-je-bestanden-veilig-in-de-cloud-met-surfdrive/veelgestelde-vragen-over-surfdrive?dst=n1469).

Awre, Chris, Jim Baxter, Brian Clifford, Janette Colclough, Andrew Cox, Nick Dods, Paul Drummond, Yvonne Fox, Martin Gill, Kerry Gregory, Anita Gurney, Juliet Harland, Masud Khokhar, Dawn Lowe, Ronan O’Beirne, Rachel Proudfoot, Hardy Schwamm, Andrew Smith, Eddy Verbaan, Liz Waller, Laurian Williamson, Martin Wolf, and Matthew Zawadzki. 2015. “Research Data Management as a ‘Wicked Problem.’” *Library Review* 64(4/5):356–71.

Borgman, Christine L. 2017. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Vol. 20. MIT press.

Bryman, Alan. 2012. *Social Research Methods*. 4th ed. Oxford ; New York: Oxford University Press.

Burgi, Pierre-Yves, Eliane Blumer, and Basma Makhlouf-Shabou. 2017. “Research Data Management in Switzerland.” *IFLA Journal* 43(1):5–21.

Buys, Cunera M., and Pamela L. Shaw. 2015. “Data Management Practices Across an Institution: Survey and Report.” *Journal of Librarianship & Scholarly Communication* 3(2):1–24.

Cox, Andrew M., and Stephen Pinfield. 2014. “Research Data Management and Libraries: Current Activities and Future Priorities.” *Journal of Librarianship and Information Science* 46(4):299–316.

Cox, Andrew M., Stephen Pinfield, and Jennifer Smith. 2016. “Moving a Brick Building: UK Libraries Coping with Research Data Management as a ‘Wicked’ Problem.” *Journal of Librarianship and Information Science* 48(1):3–17.

DCMI. 2018. “DCMI: Home.” *Dcmi*. Retrieved October 10, 2018 (http://dublincore.org/).

Demchenko, Yuri, Zhiming Zhao, Paola Grosso, Adianto Wibisono, and Cees De Laat. 2012.

"Addressing Big Data Challenges for Scientific Data Infrastructure." Pp. 614–17 in *CloudCom 2012 - Proceedings: 2012 4th IEEE International Conference on Cloud Computing Technology and Science*. IEEE.

Diener, Edward, and Rick Crandall. 1978. *Ethics in Social and Behavioral Research.* U Chicago Press.

Dunning, Alastair, Annemiek Kuil van der, Madeleine Smaele de, Teperek Marta, and Versteeg Anke. 2018. *TU Delft Research Data Framework Policy*.

European Commission. 2013. "Guidelines on Data Management in Horizon 2020." (December):6.

Flores, Jodi Reeves, Jason J. Brodeur, Morgan G. Daniels, Natsuko Nicholls, and Ece Turnator. 2015. "Libraries and the Research Data Management Landscape." *The Process of Discovery: The CLIR Postdoctoral Fellowship Program and the Future of the Academy* 82–102.

Friese, Susanne. 2017. "How to Make the Best of Codes in ATLAS.Ti." *Atlas.Ti* 1–5.

Gregor, Shirley, and Alan R. Hevner. 2013. "Positioning and Presenting Design Science Research for Maximum Impact." *MIS Quarterly* 37(2):337–55.

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004. "Design Science in Information Systems Research." *MIS Quaterly* 28(4):75–105.

Jones, Sarah. 2013. "Bringing It All Together : Research Data Management at Monash." *DCC RDM Services Case Studies* (March):1–8.

Jones, Sarah, Graham Pryor, and Angus Whyte. 2013. "A Digital Curation Centre 'working Level' Guide How to Develop Research Data Management Services -a Guide for HEIs." *DCC How-to Guides*.

Kissel, Richard, Gary Locke, and Patrick D. Gallagher. 2011. *Glossary of Key Information Security Terms?*

Klindt, Marco, and Kilian Amrhein. 2015. "One Core Preservation System for All Your Data. No Exceptions!" *IPRES* 101.

Lewis, John, and A. 2014. *Research Data Management Technical Infrastructure : A Review of Options for Development at the University of Sheffield*.

Lewis, M. J. 2010. "Libraries and the Management of Research Data." *Envisioning Future Academic Library Services* 145–68.

Mark Saunders, Philip Lewis, Adrian Thornhill. 2015. *Methoden En Technieken van Onderzoek*. 7th ed. Pearson Benelux B.V.

Marshall, Catherine, and Gretchen B. Rossman. 2006. *Designing Qualitative Research*. Vol. 9. 4th ed.

National Science Board, and National Science Foundation. 2005. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." *National Science Board* 87.

NFU. n.d. "NFU Data4lifesciences | Handbook for Adequate Natural Data Stewardship." Retrieved December 11, 2017 (http://data4lifesciences.nl/hands/handbook-for-adequate-natural-data-stewardship/).

Oleksik, Gerard, Natasa Milic-Frayling, and Rachel Jones. 2014. "Study of Electronic Lab Notebook Design and Practices That Emerged in a Collaborative Scientific Environment." Pp. 120–33 in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. New York, New York, USA: ACM Press.

Parsons, Thomas, Shirley Grimshaw, and Laurian Williamson. 2013. *University of Nottingham Research Data Management Survey*.

Pilat, Dirk, and Yukiko Fukasaku. 2007. "OECD Principles and Guidelines for Access to Research Data from Public Funding." *Data Science Journal* 6:OD4–11.

Rebouillat, Violaine. 2017. "Inventory of Research Data Management Services in France." Pp. 174–81 in *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices - Proceedings of the 21st International Conference on Electronic Publishing*.

Rice, Robin, Çuna Ekmekcioglu, Jeff Haywood, Sarah Jones, Stuart Lewis, Stuart Macdonald, and Tony Weir. 2013. "Implementing the Research Data Management Policy: University of Edinburgh Roadmap." *International Journal of Digital Curation* 8(2):194–204.

Rubacha, Michael, Anil K. Rattan, and Stephen C. Hosselet. 2011. "A Review of Electronic Laboratory Notebooks Available in the Market Today." *Journal of Laboratory Automation* 16(1):90–98.

Sewerin, Cristina, Dylanne Dearborn, Angela Henshilwood, and Michelle Spence. 2015. "Research Data Management Faculty Practices: A Canadian Perspective." *Proceedings of the IATUL Conferences* Paper 2.

Silver, Christina, and Nicholas H. Woolf. 2018. *Qualitative Analysis Using ATLAS. Ti: The Five-Level QDA Method*. Routledge.

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe Hourclé, John Ernest Kratz, Jennifer Lin, Lars Holm Nielsen, Amy Nurnberger, Stefan Proell, Andreas Rauber, Simone Sacchi, Arthur Smith, Mike Taylor, and Tim Clark. 2015. "Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications." *PeerJ Computer Science* 1:e1.

Tremblay, M. C., A. R. Hevner, and D. J. Berndt. 2008. "The Use of Focus Groups in Design Science Research." in *Proceedings of the 3rd International Conference on Design Science Research in Information Systems and Technology, DESRIST 2008*.

Truex, Duane, Steven Alter, and Cherie Long. 2010. "Empowering Business Professionals through a Systems Analysis Method That Fits Their Needs." *ECIS 2010 Proceedings* 1–12.

UKDataService. 2016. "UK Data Service." Retrieved December 11, 2017 (https://www.ukdataservice.ac.uk/).

Universiteit Leiden. n.d. "Datamanagement." Retrieved December 19, 2017 (https://www.bibliotheek.universiteitleiden.nl/onderzoek-en-publiceren/datamanagement).

Verhaar, Peter, Fieke Schoots, Laurents Sesink, and Floor Frederiks. 2017. "Fostering Effective Data Management Practices at Leiden University." *LIBER Quarterly* 27(1):1–22.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. .. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3:160018.

Wittenberg, Jamie, and Mary Elings. 2017. “Building a Research Data Management Service at the University of California, Berkeley.” *IFLA Journal* 43(1):89–97.

Yousuf, Muhammad Imran. 2007. “Using Experts`Experts`Opinions Through Delphi Technique.” *Practical Assessment, Research, and Evaluation* 12(4).

7. Bijlage literatuurlijst voor literatuuronderzoek

Tabel met literatuur

Uiteindelijke literatuurlijst die op inhoud en context is gecontroleerd								
Cites	Authors	Title	Year	Source	Publisher	Type	Gebruikt ?	Opmerkingen
71	Akers KG, Doty J	Disciplinary differences in faculty research data management practices and perspectives	2013	International Journal of Digital Curation	ijdc.net	JOUR	Ja	
18	Alexogiannopoulos E, McKenney S, Pickton M	Research Data Management Project: a DAF investigation of research data management practices at The University of Northampton	2010		nectar.northampton.ac.uk	RPRT	Ja	
340	Alter S	The work system method: connecting people, processes, and IT for business results	2006		books.google.com	BOOK	Ja	
324	Alter S	Defining information systems as work systems: implications for the IS field	2008	European Journal of Information Systems	Springer	JOUR	Ja	
	Alter S	Work systems and service systems … Where technology contributes to business results	n.d.	Stevenalter.com	www.stevenalter.com	WEB	Ja	
11	Awre C, Baxter J, Clifford B, Colclough J, Cox AM…	Research data management as a “wicked problem”	2015	Library review	emeraldinsight.com	JOUR	Ja	
268	Borgman CL	Big data, little data, no data: Scholarship in the networked world	2015		MIT press	BOOK	Ja	
4	Burgi PY, Blumer E, Makhlouf-Shabou B	Research data management in Switzerland: National efforts to guarantee the sustainability of research outputs	2017	IFLA journal	journals.sagepub.com	JOUR	Ja	
13	Buys CM, Shaw PL	Data Management Practices Across an Institution: Survey and Report.	2015	Journal of Librarianship & Scholarly Communication	jlsc-pub.org	RPRT	Ja	
105	Cox AM, Pinfield S	Research data management and libraries: Current activities and future priorities	2014	Journal of Librarianship and Information Science	journals.sagepub.com	JOUR	Ja	
20	Cox AM, Pinfield S, Smith J	Moving a brick building: UK libraries coping with research data management as a 'wicked'problem	2016	Journal of Librarianship and …	journals.sagepub.com	JOUR	Ja	
	DCMI	Dublin Core Metadata Initiative	2018	Website van DCMI	http://dublincore.org	WEB	Ja	
0	Delden P van	Wicked problems: Venijnige vraagstukken als vuurproef voor de publieke dienstverlening	2014			ICOMM	Ja	Alleen voor vertaling Wicked problem
101	Demchenko Y, Zhao Z, Grosso P…	Addressing big data challenges for scientific data infrastructure	2012	In Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on (pp. 614-617)	ieeexplore.ieee.org	CONF	Ja	Zijdelings gebruikt, komt uit eerder vak
2	European commission	Guidelines on FAIR Data Management in Horizon 2020	2016	http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf		WEB	Ja	
12	Flores JR, Brodeur JJ, Daniels MG…	Libraries and the research data management landscape	2015	The process of discovery: The CLIR postdoctoral fellowship program and the future of the academy	Council on Library and Information Resources	RPRT	Ja	

Cites	Authors	Title	Year	Source	Publisher	Type	Gebruikt ?	Opmerkingen
1004	Gregor S, Hevner AR	Positioning and presenting design science research for maximum impact.	2013	MIS quarterly	ai.arizona.edu	JOUR	Ja	
10025	Hevner AR, March ST, Park J, Ram S	Design science in information systems research	2004	MIS quarterly	Springer	JOUR	Ja	
2	Jones S	Bringing it all together: a case study on the improvement of research data management at Monash University	2013	dcc website	www.dcc.ac.uk	WEB	Ja	
32	Jones S, Pryor G, Whyte A	How to Develop Research Data Management Services-a guide for HEIs.	2013	dcc website	www.dcc.ac.uk	WEB	Ja	
3	Klindt M, Amrhein K	One Core Preservation System for All your Data. No Exceptions!	2015	PRES 2015 - Proceedings of the 12th International Conference on Preservation of Digital Objects	opus4.kobv.de	CONF	Ja	doorverwijzing
4	Lewis JA	Research data management technical infrastructure: A review of options for development at the University of Sheffield	2014		wiki.lib.sun.ac.za	RPRT	Ja	Was wel gevonden bij ELN
48	Lewis MJ	Libraries and the management of research data.	2010	Envisioning Future Academic Library Services	http://www.facetpublishing.co.uk	BOOK	Ja	Zijdelings gebruikt.
21	National Science Board	Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century	2005	National Science Board, & National Science Foundation	https://www.nsf.gov/nsb/	WEB	Ja	
	NFU	NFU Data4lifesciences | Handbook for Adequate Natural Data Stewardship	n.d.	NFU	rom http://data4lifesciences.nl/hands/handbook-for-adequate-natural-data-stewardship	WEB	Ja	
10	Oleksik, G, Milic-Frayling, N, & Jones, R	Study of electronic lab notebook design and practices that emerged in a collaborative scientific environment.	2014	Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing	dl.acm.org	CONF	Ja	Zijdelings gebruikt, komt uit eerder vak
12	Parsons T, Grimshaw S, Williamson L	Research data management survey: report	2013		eprints.nottingham.ac.uk	RPRT	Ja	
31	Pilat D, Fukasaku Y	OECD principles and guidelines for access to research data from public funding	2007	Data Science Journal	jstage.jst.go.jp	JOUR	Ja	
0	Rebouillat V	Inventory of research data management services in France	2017	Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices - Proceedings of the 21st International Conference on Electronic Publishing	http://www.cyprusconferences.org/elpub2017/	CONF	Ja	
12	Rice R, Ekmekcioglu Ç, Haywood J, Jones S…	Implementing the research data management policy: University of Edinburgh roadmap	2013	International Journal of Digital Curation	ijdc.net	JOUR	Ja	
11418	Rittel HWJ, Webber MM	Dilemmas in a general theory of planning	1973	Policy sciences	Springer	JOUR	Ja	Doorverwijzing
65	Rubacha, M, Rattan, AK, & Hosselet, SC	A Review of Electronic Laboratory Notebooks Available in the Market Today		SLAS TECHNOLOGY: Translating Life Sciences Innovation	Journals.sagepub.com	JOUR	Ja	Zijdelings gebruikt, komt uit eerder vak
1	Sewerin S	Research data management faculty practices: A Canadian perspective	2015	2015 IATUL Proceedings	docs.lib.purdue.edu	CONF	Ja	
55	Starr J, Castro E, Crosas M, Dumontier M	Achieving human and machine accessibility of cited data in scholarly publications	2015	PeerJ computer Science	peerj.com	JOUR	Ja	
50	Truex D, Alter S, Long C	Systems analysis for everyone else: Empowering business professionals through a systems	2010	ECIS	diva-portal.org	CONF	Ja	

Cites	Authors	Title	Year	Source	Publisher	Type	Gebruikt ?	Opmerkingen
		analysis method that fits their needs						
	TU Delft Research Support	RS Portal: Make your Data Management Plan	n.d.	Researchsupport website	http://researchsupport.tudelft.nl/experimentation/manage-archive-your-research-data/make-your-data-management-plan/	WEB	Ja	
	UKDataService	UK Data Service	2016	Website van UK dataservice	https://www.ukdataservice.ac.uk/	WEB	Ja	
	Universiteit Leiden	Datamanagement	n.d.	Bibliotheek Universiteit Leiden, datamanagement	https://www.bibliotheek.universiteitleiden.nl/onderzoek-en-publiceren/datamanagement	WEB	Ja	
2	Verhaar P, Schoots F, Sesink L, Frederiks F	Fostering effective data management practices at Leiden University	2017	Liber Quarterly	liberquarterly.eu	JOUR	Ja	
432	Wilkinson MD, Dumontier M, Aalbersberg IJJ…	The FAIR Guiding Principles for scientific data management and stewardship	2016	Scientific data	nature.com	JOUR	Ja	
485	Borgman CL	The conundrum of sharing research data	2012	Journal of the Association for Information …	Wiley Online Library	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
48	Carlson J	Demystifying the data interview: developing a foundation for reference librarians to talk with researchers about their data	2012	Reference Services Review	emeraldinsight.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
23	Chard k, Pruyne J, Blaiszik B…	Globus data publication as a service: Lowering barriers to reproducible science	2015	2015 IEEE 11th International Conference on e-Science	ieeexplore.ieee.org	CONF	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
2	Clement R, Blau A, Abbaspour P…	Team-based data management instruction at small liberal arts colleges	2017	IFLA Journal	journals.sagepub.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
19	Clements A, McCutcheon V	Research data meets research information management: Two case studies using (a) Pure CERIF-CRIS and (b) EPrints repository platform with CERIF extensions	2014	Elsevier, Procedia Computer Science	www.sciencedirect.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
19	Davidson J, Jones S, Molloy L, Kejser UB	Emerging good practice in managing research data and research information within UK Universities	2014	Elsevier, Procedia Computer Science	www.sciencedirect.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
7	Frey JG, Milsted A, Michaelides D…	MyExperimentalScience, extending the 'workflow'	2013	Concurrency and Computation: Practice and Experience	Wiley Online Library	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
0	Gruetz R, Franke T, Dickmann F	Concept for preservation and reuse of genome and biomedical imaging research data.	2013	Studies in health technology and …	europepmc.org	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data

Cites	Authors	Title	Year	Source	Publisher	Type	Gebruikt ?	Opmerkingen
14	Henderson ME, Knott TL	Starting a research data management program based in a university library	2015	Medical reference services quarterly	Taylor & Francis www.tandfonline.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
9	Hiom D, Fripp D, Gray S, Snow K…	Research data management at the University of Bristol: charting a course from project to service	2015	Program: electronic library and information systems	emeraldinsight.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
4	Knight G	Building a research data management service for the London School of Hygiene & Tropical Medicine	2015	Program: electronic library and information systems	emeraldinsight.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
79	Lyon L	The informatics transform: Re-engineering libraries for the data decade	2012	International Journal of Digital Curation	ijdc.net	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
74	Pryor G, Donnelly M	Skilling up to do data: whose role, whose responsibility, whose career?	2009	International Journal of Digital Curation	ijdc.net	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
25	Richardson J, Nolan-Brown T, Loria P…	Library research support in Queensland: a survey	2012	Taylor and Francis	www.tandfonline.com			Onvoldoende inhoudelijke aansluiting bij managing active data
65	Scaramozzino JM, Ramírez ML…	A study of faculty data curation behaviors and attitudes at a teaching-centered university	2012	College & Research libraries	digitalcommons.calpoly.edu	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
4	Siart C, Kopp S, Apel J	The Interface between Data Science, Research Assessment and Science Support-Highlights from the German Perspective and Examples from Heidelberg University	2015	Advanced Applied Informatics (IIAI-AAI)	ieeexplore.ieee.org	CONF	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
2	Simons E, Jetten M, Messelink MC, Berchem M…	The Important Role of CRIS's for Registering and Archiving Research Data. The RDS-project at Radboud University (the Netherlands) in Cooperation with Data …	2017	Elsevier, Procedia computer science	www.sciencedirect.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
605	Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M & Frame M	Data sharing by scientists: practices and perceptions	2011	PloS one	journals.plos.org	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij managing active data
4	Wittenberg J, Elings M	Building a research data management service at the University of California, Berkeley: a tale of collaboration	2017	IFLA journal	journals.sagepub.com	JOUR	Nee	Onvoldoende inhoudelijke aansluiting bij

Cites	Authors	Title	Year	Source	Publisher	Type	Gebruikt ?	Opmerkingen
								managing active data
Opmerkingen: Cites gebaseerd op Google Scholar februari 2018								

Tabel 24, literatuurlijst met laatste inhoudelijke keuze

8. Bijlage ontwikkeling van de probleemstelling

De eerste probleemstelling wordt gedurende het onderzoek steeds verder aangepast. Initieel is het idee om ELN's te gebruiken om data vindbaar en volgbaar te houden (idee gebaseerd op het artikel van Oleksik et al (2014)).

Aan welke principes moeten ELN's en onderzoeksdata-opslag voldoen om ze geschikt te maken als middel voor onderzoekers om alle relevante onderzoeksdata van een onderzoek vindbaar, volgbaar en op eenvoudige wijze archiveerbaar te maken binnen een universiteit?

Er zijn niet veel artikelen die de combinatie ELN en RDM bevatten. Bovendien blijkt uit de surveys dat lang niet alle wetenschappers een ELN gebruiken (Parsons et al. 2013; Sewerin et al. 2015). Het gebrek aan gevonden bruikbare artikelen en mijn dagelijkse werkgebied (Systems en Storage) doen mij in overleg met mijn begeleider besluiten, de probleemstelling aan te passen:

Aan welke ontwerpprincipes moet een Research Data Management werksysteem, toegespitst op het beheer van actieve onderzoeksdata, binnen een academische omgeving, voldoen zodat deze data vindbaar, volgbaar en archiveerbaar zijn?

Na de literatuurstudie, wordt het WSS als design artifact opgeleverd. In de empirische fase zal dan een nieuw WSS, gebaseerd op het empirisch onderzoek, geconfronteerd worden met de architectuurprincipes van de case organisatie. Het gevonden WSS na literatuuronderzoek is groot. Het is een goed globaal ontwerp. Om de scope van het onderzoek beperkt te houden, wordt besloten om af te zien van de confrontatie van het WSS met de architectuurprincipes. Hiermee wordt nog steeds voldaan aan de voorwaarden voor DSR, waarbij een design artifact in de vorm van een model wordt opgeleverd.

Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op het beheer van actieve onderzoeksdata, binnen de case organisatie, voldoen zodat deze data vindbaar, volgbaar en archiveerbaar zijn?

Tijdens de interviews worden drie zaken duidelijk:

- Om (delen van) onderzoek te herhalen zijn behalve onderzoeksdata ook metadata en datadocumentatie nodig;
- De enige persoon die kan bepalen welke data hiervoor relevant zijn, is de PI;
- De eerste gedachte was, dat actieve data data waren die tijdens het lopende onderzoek werden verzameld en/of bewerkt. In de interviews bleek dat data ook als actief kunnen worden beschouwd na het onderzoek. De formulering wordt aangepast naar data die verzameld en bewerkt is tijdens het lopende onderzoek.

Het begrip relevante data voor lopend onderzoek wordt gelanceerd (onderzoeksdata, metadata en datadocumentatie die volgens de onderzoeker tezamen nodig zijn om delen van het onderzoek te kunnen herhalen).

Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op de relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en archiveerbaar zijn?

De data opslag die wordt gebruikt voor archivering is niet hetzelfde als de opslag die men gebruikt tijdens het onderzoek. Verder blijkt dat men bij voorkeur de relevante data zo lang mogelijk laat staan op de locatie die tijdens het onderzoek gebruikt is. Het gaat dus niet om archivering met een PID en een door het vakgebied geaccepteerde repository, maar om langdurige opslag voor een relevante periode. Ook in dit geval kan alleen de PI bepalen hoe lang een relevante periode is.

Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op de relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en voor relevante periode op te slaan zijn?

De belangrijkste wijzigingen zijn die van actieve onderzoeksdata naar relevante data voor lopende onderzoeken en die van archivering naar opslaan voor een relevante periode.

- Actief en voor lopende onderzoeken hebben dezelfde intentie als waar mee begonnen is (het is een andere formulering). Relevante data is breder dan onderzoeksdata. Het bevat ook de metadata en de datadocumentatie. Daarentegen is het ook wat beperkter, aangezien het alleen gaat om de data die nodig zijn om (delen van) het onderzoek te herhalen. Er is geen informatie verloren gegaan door de overstap van onderzoeksdata naar relevante data. Het is een leerpunt uit de interviews. Van begin af aan werd in de interviews aan alle datavormen aandacht besteed en was ook snel duidelijk dat er meer nodig was dan onderzoeksdata om onderzoek te kunnen herhalen.
- Archivering naar opslaan voor een relevante periode is ook iets wat zich tijdens de interviews heeft ontwikkeld. Tijdens het onderzoek is men beperkt bezig met de toekomstige archivering. Men wil vooral voldoende opslag hebben om voor een periode die relevant is voor hen, de data te kunnen laten staan. Ook dit is een leerpunt uit de interviews.

9. Bijlage samenvoeging van de resultaten voor products and services en technology t.a.v. vindbaarheid, volgbaarheid en archiveerbaarheid

Vanuit verschillende invalshoeken is onderzocht in hoeverre resultaten uit de literatuur vindbare, volgbare en/of archiveerbare eigenschappen van het werksysteem verwoorden (stellingen). Er is daarbij gekeken naar dataprincipes (FAIR en OECD), naar 'products and services' en naar 'technology'. In deze bijlage worden per aspect (vindbaarheid, volgbaarheid, archiveerbaarheid) de dataprincipes naast de gevonden 'products and services' en stellingen gelegd en naast de gevonden 'technology' stellingen gelegd. De stellingen en principes worden (waar mogelijk en waar logisch) samengevoegd. Daar waar de stellingen en principes elkaar in betekenis overlappen, worden ze ontdubbeld. Omdat er wordt begonnen met stellingen, worden de resultaten (om het onderscheid te kunnen bewaren) bevindingen genoemd.

Hieronder staan de resultaattabellen herhaald. Alleen de stellingen met een positief resultaat voor vindbaarheid, volgbaarheid en archiveerbaarheid worden getoond. In de tabel voor dataprincipes zijn de titelrijen weggehaald en de resultaten (net als bij de tabel voor 'products and services') op volgorde gezet.

FAIR en OECD principes toegepast op de actieve fase met de aspecten vind-, volg- en archiveerbaar				
Van toepassing zijnde aspecten probleemstelling	*VDB[1]*	*VGB[2]*	*ARB[3]*	Commentaar
(meta)data zijn vindbaar in een doorzoekbare bron	X			Duidelijk zonder verdere uitleg
Openness; toegang voor de internationale research gemeenschap	X			De gemeenschap moet daarvoor kunnen vinden in de actieve fase.
Data worden met rijke metadata beschreven	X	X		Volgt definitie metadata.
(meta)data gebruiken een formele, toegankelijke, gedeelde en wijds beschikbare taal voor kennisrepresentatie	X	X		Volgt definitie metadata.
(meta)data gebruiken vocabulaires die de FAIR principes volgen	X	X		Volgt definitie metadata.
(meta)data zijn rijk beschreven met voldoende relevante attributen	X	X		Volgt definitie metadata.
Security; het waarborgen van de veiligheid en integriteit van de data bij het verlenen van toegang aan derden.		X	X	Kunnen zien dat de oorspronkelijke dataset onveranderd is, geldt voor en na de actieve fase.
(meta)data worden met een duidelijke en toegankelijke dataovereenkomst beschikbaar gesteld			X	Dit geldt voor te archiveren data.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid [3]ARB: Archiveerbaarheid				

Tabel 25, dataprincipes voor het vereenvoudigen en samenvoegen van resultaten

Products and services afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid				
Product en/of service	VDB[1]	VGB[2]	ARB[3]	Commentaar
Het werksysteem verzekert dat er gedurende het onderzoek voldoende betrouwbare opslagruimte beschikbaar is om de data op te slaan (Cox and Pinfield 2014; Demchenko et al. 2012; Jones et al. 2013; Lewis and A. 2014; Rice et al. 2013; Sewerin et al. 2015) (Repository selection uit TABEL 11);	X			Bekende locatie maakt vinden makkelijker.
Het werksysteem levert sync and share functionaliteit ('academic dropbox') (Alexogiannopoulos et al. 2010; Buys and Shaw 2015; Jones et al. 2013; Lewis and A. 2014; Parsons et al. 2013; Rice et al. 2013);	X			Het syncen naar één centrale locatie verhoogt de vindbaarheid van de mogelijk verspreide data.
Het werksysteem biedt de mogelijkheid dat data van electronic lab notebooks (ELN) verplaatst worden naar centrale active data opslag (Lewis and A. 2014).	X			Als de data op 1 locatie staan, wordt vindbaarheid vereenvoudigd.
Het werksysteem maakt het mogelijk om de data vanaf elk device (pc's, laptops, tablets, telefoons, laboratoriuminstrumenten enz.) te benaderen (Lewis and A. 2014; Rice et al. 2013);	X			Als de data op 1 locatie staan, wordt vindbaarheid vereenvoudigd.
Het werksysteem biedt de mogelijkheid om data naar meerdere locaties/systemen te distribueren/repliceren. Rice et al. noemen specifiek HPC omgevingen (Demchenko et al. 2012; Rice et al. 2013);	X			Op voorwaarde van geschikte logging blijven de data vindbaar.
Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden (Verhaar et al. 2017).	X			De structuur vereenvoudigt vinden.
Het werksysteem verzekert dat onderzoeksdata vindbaar en begrijpelijk zijn door de data in combinatie met de bijbehorende metadata en andere documentatie op te slaan (Cox and Pinfield 2014; Lewis and A. 2014; Sewerin et al. 2015; Verhaar et al. 2017)(data documentation in TABEL 11);	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
Het werksysteem verzekert dat backupmogelijkheden gedurende het onderzoek beschikbaar zijn (Lewis and A. 2014; Sewerin et al. 2015; Verhaar et al. 2017);	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.
Het werksysteem zorgt dat het aanmaken van metadata bij voorkeur geautomatiseerd plaatsvindt (Jones et al. 2013; Lewis and A. 2014);	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en voor machines (bijvoorbeeld een API). Dit is noodzakelijk om met het werksysteem te kunnen werken.	X	X		Om data te kunnen vinden en volgen is een interface om mee te zoeken noodzakelijk.
Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen (Alexogiannopoulos et al. 2010; Jones et al. 2013; Lewis and A. 2014; Verhaar et al. 2017);		X		Helpt datatransformaties inzichtelijk te maken.
Het werksysteem levert de mogelijkheid om data te vernietigen (Lewis and A. 2014);		X		Het is de laatste bewerking op de dataset. Voorwaarde is voldoende logging.
Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan (Sewerin et al. 2015; Verhaar et al. 2017);		X		Op voorwaarde van geschikte logging volgbaar.
Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering) (Lewis and A. 2014);		X	X	Volgbare transformatie en klaarmaken voor archief.
Het werksysteem kan unieke identifiers aanmaken voor een dataset (Persistent Identifiers, PID) (Lewis and A. 2014).			X	Een PID* is nodig voor archiveerbaarheid.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid				

Product en/of service	VDB[1]	VGB[2]	ARB[3]	Commentaar
[3]ARB: Archiveerbaarheid *Een PID is een persistent identifier, een unieke code waarmee de dataset geïdentificeerd kan worden.				

Tabel 26, 'products and services' voor het vereenvoudigen en samenvoegen van resultaten

RDM infrastructuur voorwaarden afgezet tegen vindbaarheid, volgbaarheid en archiveerbaarheid				
RDM infrastructuur voorwaarde	VDB[1]	VGB[2]	ARB[3]	Kort Commentaar
Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface (Lewis and A. 2014).	X			Op voorwaarde dat de register/index voor doorzoekbaarheid zorgt, een vindbaarheid stelling.
Het werksysteem biedt opslag voor de RDM infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd) (Burgi et al. 2017).	X			Fysieke vindbaarheid, weten waar de data daadwerkelijk staan.
Opmerkingen: [1]VDB: Vindbaarheid [2]VGB: Volgbaarheid [3]ARB: Archiveerbaarheid				

Tabel 27, 'technology' voor het vereenvoudigen en samenvoegen van resultaten

Op basis van deze tabellen worden in de volgende paragrafen bevindingen voor vindbaarheid, vindbaarheid en volgbaarheid, volgbaarheid, volgbaarheid en archiveerbaarheid, archiveerbaarheid en vindbaarheid, volgbaarheid en archiveerbaarheid afgeleid voor ‘products and services’ en voor ‘technology’. .

De dataprincipes uit Tabel 25 maken blauwe rijen in de nieuwe tabellen.

De ‘products and services’ resultaten uit Tabel 26 maken okere rijen in de nieuwe tabellen.

De ‘technology’ resultaten uit Tabel 27 maakt groene rijen in de nieuwe tabel in de volgende paragraaf.

In veel van de stellingen worden metadata genoemd. Metadata hebben belangrijke eigenschappen als het gaat om vindbaarheid en volgbaarheid. Daarom wordt eerst de definitie van metadata in deze bijlage herhaald.

Metadata: Metadata zijn data die informatie geven over de onderzoeksdata met als doel de onderzoeksdata voor anderen bruikbaar te maken (reproduceerbaar en interpreteerbaar) (National Science Board and National Science Foundation 2005). Metadata kunnen in lagen geformuleerd worden. De eerste laag is een breed toepasbare standaard (zoals de Dublin Core standaard voor metadata (DCMI 2018)). Deze wordt gevolgd door een domein/discipline specifieke laag en vervolledigd met een laag die de data op folder-, file- en naamniveau beschrijft (Wilkinson et al. 2016). Metadata zouden tenminste moeten bestaan uit een dataset identifier, titel, beschrijving, de maker, contact (persoon/instituut), publicatiedatum, versie, identifiers van de makers en de licentie (Starr et al. 2015). Dataset identifiers, de maker, het contact en identifiers van de makers helpen bij de vindbaarheid van de data. Publicatiedatum, beschrijving en versie helpen bij de volgbaarheid van data.

Enkele andere belangrijke begrippen die van belang zijn bij het samenvoegen van bepaalde stellingen zijn:

Synchroniseren: Data naar een tweede (of meer dan dat) locatie repliceren. Dat kan een eenmalige actie zijn (een kopie), het kan realtime zijn, waarbij de beide omgevingen voortdurend gelijk zijn (spiegelen) of dat data op gezette tijden wordt gerepliceerd.

Onderzoeksdatahouders: Alle decentrale devices en of systemen (en mogelijk hun backups naar lokale media) waar onderzoeksdata op verzameld worden (ELN's, HPC omgevingen, meetinstrumenten, laptops, desktops, flashdrives enzovoort).

Aangezien er alleen 'vindbare' 'technology' stellingen zijn, worden die eerst behandeld waarna de rest van de bijlage over de 'products and services' en dataprincipes per aspect gaat.

1.1. Vindbaar 'technology'

Hieronder de resultaten voor 'vindbaar' voor dataprincipes en 'technology':

dataprincipes en 'technology' voor vindbaarheid	
1	(meta)data zijn vindbaar in een doorzoekbare bron
2	Openness; toegang voor de internationale research gemeenschap
3	Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface (Lewis and A. 2014).
4	Het werksysteem biedt opslag voor de RDM infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd) (Burgi et al. 2017).

Tabel 28, dataprincipes en 'technology' voor vindbaarheid

Stelling 1 stelt dat data en metadata vindbaar dienen te zijn in een doorzoekbare bron.

Stelling 3 stelt dat er een register/index voor metadata en/of datadocumentatie moet zijn.

Een register/index kan beschouwd worden als een mogelijke technische invulling van een doorzoekbare bron. Dit kan dan samengevat worden naar de eis:

- Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.

Dit is meer een dienst dan een technische vraag. De eis verhuist daardoor naar de lijst van 'vindbare' 'products and services'.

Stelling 4 lijkt verder niet te combineren met andere stellingen. Wel is het zo dat redundantie een technische invulling van betrouwbaarheid is. Stelling 3 geeft een technische invulling van 'doorzoekbare bron'. Hierdoor blijven de twee 'technology' bevindingen onveranderd staan.

- Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface.
- Het werksysteem biedt opslag voor de RDM infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd).

1.2. Vindbaar ‘products and services’

Hieronder eerst de resultaten uit de tabellen voor dataprincipes en ‘products and services’:

dataprincipes en 'products and services' voor vindbaarheid	
1	(meta)data zijn vindbaar in een doorzoekbare bron
2	Openness; toegang voor de internationale research gemeenschap
3	Het werksysteem verzekert dat er gedurende het onderzoek voldoende betrouwbare opslagruimte beschikbaar is om de data op te slaan
4	Het werksysteem levert sync and share functionaliteit (‘academic dropbox’)
5	Het werksysteem biedt de mogelijkheid dat data van electronic lab notebooks (ELN) verplaatst worden naar centrale active data opslag
6	Het werksysteem maakt het mogelijk om de data van elk device (pc’s, laptops, tablets, telefoons, laboratoriuminstrumenten enz.) te kopiëren
7	Het werksysteem biedt de mogelijkheid om data naar meerdere locaties/systemen te distribueren/repliceren. Rice et al. noemen specifiek HPC
8	Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden

Tabel 29, dataprincipes en 'products and services' voor vindbaarheid

Stelling 3 stelt dat opslagruimte voor onderzoeksdata voldoende (capaciteit) en betrouwbaar dient te zijn.

Stelling 4 stelt dat dat er sync functionaliteit, in geval van Dropbox tussen een decentrale bron (zoals een onderzoeksdatahouder) en een centrale opslagplaats dient te zijn. (De in stelling 4 genoemde share functionaliteit wordt bij de volgende eis behandeld.)

Stelling 5 stelt dat er een mogelijkheid om data te kopiëren van ELN (onderzoeksdatahouder) naar een centrale opslagplaats dient te zijn.

Stelling 6 stelt dat er een mogelijkheid om data van diverse devices (onderzoeksdatahouders) te kopiëren dient te zijn. Er wordt hier aangenomen dat dat naar centrale opslag gaat.

Stelling 7 stelt dat het kopiëren van data (vanaf centrale opslag) naar locaties/systemen, met name HPC (onderzoeksdatahouders) moet kunnen plaatsvinden. (Stelling 7 voldoet alleen als zijnde een vindbare stelling, als er voldoende relevante logging aanwezig is (Tabel 26). Hier wordt in “Vindbaar en volgbaar” op teruggekomen).

Door de brede definitie van synchroniseren te hanteren en het begrip onderzoeksdatahouder te gebruiken, en de eisen voor betrouwbaarheid en capaciteit uit de stelling 3 toe te voegen, kan dit worden samengevat in de eis:

- Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders (zowel van als naar).

Stelling 2 stelt dat data gedeeld moeten kunnen worden met de internationale research gemeenschap.

Stelling 4 stelt dat het werksysteem share functionaliteit dient aan te bieden.

Als de ‘volgbaar en archiveerbare’ stelling om de veiligheid en integriteit van de data te waarborgen wordt meegenomen, dan kunnen stelling 2 en 4 hiermee samengevoegd worden tot de volgende eis die daarmee geldt voor vindbaar, volgbaar en archiveerbaar:

- Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.

De eis komt in paragraaf 1.7 (vindbaar, volgbaar en archiveerbaar) terug.

Het userinterface uit stelling 3 kan aangevuld worden met de 'vindbaar en volgbare' stelling dat er buiten een userinterface voor mensen ook een interface (API) voor machines moet zijn. Uit 4.9 (paragraaf over de activiteiten binnen het werksysteem) volgt dat de interfaces o.a. gebruikt kunnen worden om opdrachten te geven, resultaten te lezen en logs te bekijken. Men kan er dus ook de PID mee aanvragen.

- Het werksysteem heeft een userinterface en API om informatie te tonen, activiteiten te volgen, logs te lezen en (zoek)opdrachten te geven.

Deze eis beschrijft een dienst en is een vindbare, volgbare en archiveerbare stelling (en komt dus in die paragraaf terug).

De resulterende bevindingen voor 'products and services' voor het aspect vindbaar zijn dan:

- Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.
- Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders.
- Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.

1.3. Vindbaar en volgbaar 'products and services'

Hieronder de resultaten uit de tabellen:

dataprincipes en 'products and services' voor vindbaarheid en volgbaarheid	
1	Data worden met rijke metadata beschreven
2	(meta)data gebruiken een formele, toegankelijke, gedeelde en wijds beschikbare taal voor kennisrepresentatie
3	(meta)data gebruiken vocabulaires die de FAIR principes volgen
4	(meta)data zijn rijk beschreven met voldoende relevante attributen
5	Het werksysteem verzekert dat onderzoeksdata vindbaar en begrijpelijk zijn door de data in combinatie met de bijbehorende metadata en andere documentatie op te slaan
6	Het werksysteem verzekert dat backupmogelijkheden gedurende het onderzoek beschikbaar
7	Het werksysteem zorgt dat het aanmaken van metadata bij voorkeur geautomatiseerd plaatsvindt
8	Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en voor machines (bijvoorbeeld een API). Dit is noodzakelijk om met het werksysteem te kunnen werken.

Tabel 30, dataprincipes en 'products and services' voor vindbaarheid en volgbaarheid

Er vanuitgaande dat in stelling 6 de backupmogelijkheden net zo goed voor data als voor metadata gelden, gelden op stelling 1 en 7 na de bovenstaande stellingen allemaal voor zowel data als metadata.

Data hebben metadata is een uitgangspunt. Dat de metadata een rijke beschrijving geven, komt overeen met de stellingen 2 en 4. Metadata helpen data begrijpelijk te maken (stelling 5).

Als daar ook de FAIR eis (stelling 3) aan wordt toegevoegd, resulteert dat in de volgende eis:

- Het werksysteem beschrijft (meta)data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR principes volgen.

Een extra voorwaarde die niet uit de literatuur komt, maar een gevolg is van de stellingen 7 uit de vorige en 2 en 3 uit de volgende paragraaf gaat over de logging van de activiteiten in het werksysteem. (Het kunnen volgen van (al dan niet geoorloofde) wijzigingen aan data (stelling 3) het kunnen volgen van verwijderen van data (stelling 2) en het kunnen vinden van verplaatste data (stelling 7) zie ook Tabel 26). De extra eis voor deze paragraaf wordt daarmee:

- Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.

Stelling 6 wordt aangepast naar restoremogelijkheden (de researcher zal uiteindelijk willen kunnen 'restoren' als er een probleem is) en de formulering wordt aangepast naar een eis aan het werksysteem. Stelling 7 wordt één op één overgenomen, waarbij de formulering wordt aangepast naar een eis aan het werksysteem. De bevindingen voor vindbaar en volgbaar worden:

- Het werksysteem beschrijft (meta)data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR principes volgen.
- Het werksysteem ondersteunt het aanmaken van metadata (bij voorkeur geautomatiseerd).
- Het werksysteem verzekert dat restoremogelijkheden gedurende het onderzoek beschikbaar zijn.
- Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.

1.4. Volgbaar 'products and services'

Hieronder de resultaten uit de tabellen (alleen 'products and services')

'products and services' voor volgbaarheid	
1	Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen (Alexogiannopoulos et al. 2010; Jones et al. 2013; Lewis and A. 2014; Verhaar et al. 2017);
2	Het werksysteem levert de mogelijkheid om data te vernietigen (Lewis and A. 2014);
3	Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan (Sewerin et al. 2015; Verhaar et al. 2017);

Tabel 31, 'products and services' voor volgbaarheid

Een belangrijke extra opmerking bij stelling 2 en 3, zoals in paragraaf 1.4 geformuleerd, is dat dit volgbaarheidsstellingen zijn op voorwaarde van ter zake doende logging door het werksysteem. Die voorwaarde van logging is al geformuleerd in de laatste stelling van 1.4.

Stelling 2 is al volledig opgenomen in de vierde eis van de vorige paragraaf. Stelling 3 wordt niet volledig overgenomen in de vierde eis van de vorige paragraaf. Zij wordt hier om die reden als eis benoemd. De resulterende bevindingen zijn daarmee:

- Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen
- Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.

1.5. Volgbaar en archiveerbaar ‘products and services’

Hieronder de resultaten uit de tabellen:

dataprincipes en 'products and services' voor volgbaarheid en archiveerbaarheid	
1	Security; het waarborgen van de veiligheid en integriteit van de data bij het verlenen van toegang aan derden.
2	Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering) (Lewis and A. 2014);

Tabel 32, dataprincipes en 'products and services' voor volgbaarheid en archiveerbaarheid

Ook deze stellingen lijken niet samenvoegbaar. De eerste stelling wordt wel gebruikt in combinatie met de stelling om data internationaal beschikbaar te stellen uit de eerste paragraaf en komt daarmee terug als een eis t.a.v. vindbaarheid, volgbaarheid en archiveerbaarheid. De enige eis die dan overblijft voor deze paragraaf is:

- Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering).

1.6. Archiveerbaar ‘products and services’

Hieronder de resultaten uit de tabellen:

dataprincipes en 'products and services' voor archiveerbaarheid	
1	(meta)data worden met een duidelijke en toegankelijke dataovereenkomst beschikbaar gesteld
2	Het werksysteem kan unieke identifiers aanmaken voor een dataset (Persistent Identifiers, PID) (Lewis and A. 2014).

Tabel 33, dataprincipes en 'products and services' voor archiveerbaarheid

Een PID is een middel om gearchiveerde data (en dus niet actieve data) te vinden. Vandaar dat het niet onder vindbaarheid terugkomt.

Vanuit paragraaf 2.5.1 volgt de toevoeging dat de wetenschapper bepaalt wat relevante onderzoeksdata zijn. De stellingen lijken verder niet samenvoegbaar. Daaruit volgen de twee bevindingen:

- Het werksysteem stelt (meta)data met een duidelijke en toegankelijke dataovereenkomst beschikbaar.
- Het werksysteem kan unieke identifiers aanmaken voor relevante onderzoeksdata (Persistent Identifiers, PID). De wetenschapper bepaalt hierbij wat relevante onderzoeksdata zijn.

1.7. Vindbaar, volgbaar en archiveerbaar ‘products and services’

Er zijn in de tabellen geen stellingen die alle drie deze aspecten raken. Door verschillende stellingen samen te voegen, zoals benoemd in de paragraaf over vindbaarheid (0), resulteren er twee bevindingen:

- Het werksysteem heeft een userinterface en API om informatie te tonen, activiteiten te volgen, logs te lezen en (zoek)opdrachten te geven.
- Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.

1.8. De bevindingen op een rij

Products and services zijn zaken waar een klant om kan vragen, wat door een klant gebruikt wordt. Technology en infrastructure zijn (noodzakelijke) technische middelen die het uitvoeren van een dienst of het leveren van een product mogelijk maken. Het zijn verschillende onderdelen in het WSS. Het levert de volgende input aan 'products and services' voor het WSS op:

Vindbaar

- Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.
- Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders.
- Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.

Vindbaar en volgbaar

- Het werksysteem beschrijft (meta)data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR principes volgen.
- Het werksysteem ondersteunt het aanmaken van metadata (bij voorkeur geautomatiseerd).
- Het werksysteem verzekert dat restoremogelijkheden gedurende het onderzoek beschikbaar zijn.
- Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.

Volgbaar

- Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen
- Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.

Volgbaar en archiveerbaar

- Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering).

Archiveerbaar

- Het werksysteem stelt (meta)data met een duidelijke en toegankelijke dataovereenkomst beschikbaar.
- Het werksysteem kan unieke identifiers aanmaken voor relevante onderzoeksdata (Persistent Identifiers, PID). De wetenschapper bepaalt hierbij wat relevante onderzoeksdata zijn.

Vindbaar, volgbaar en archiveerbaar

- Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.

- Het werksysteem heeft een userinterface en API om informatie te tonen, activiteiten te volgen, logs te lezen en (zoek)opdrachten te geven.

De twee ‘technology’ bevindingen zijn:

- Het werksysteem biedt in ieder geval een opslagplaats voor bestanden (filestore), een register/index (data registry) voor metadata en/of beschrijvende documentatie en een userinterface.
- Het werksysteem biedt opslag voor de RDM-infrastructuur die redundant is (meervoudig gespiegeld over geografisch gespreide locaties uitgevoerd).

10. Bijlage Interviews

In deze bijlage staan alle voorbereidingen voor de interviews beschreven.

10.1. Selectie geïnterviewden

De stakeholders uit de literatuurstudie zijn:

1. Research communities (onderzoeksgemeenschappen, vaak ook over instituten heen);
2. Researchers (onderzoekers en master studenten);
3. Research coordinators (onderzoekscoördinatoren);
4. Federated groups (samenwerkingsverbanden van (onderdelen van) verschillende instituten om een (deel van een) RDM dienst aan te bieden);
5. Offices of research;
6. Faculties and departments (faculteiten en afdelingen van faculteiten);
7. Research libraries;
8. Campus IT services;
9. Infrastructure providers (mogelijk samenwerkende aanbieders van technische infrastructuur voor RDM diensten);
10. Computational stakeholders.

Als dit wordt vertaald naar de case organisatie geldt:

- De eerste vijf stakeholders worden binnen de TU vertegenwoordigd door onderzoekers en onderzoekscoördinatoren (bij de case organisatie de PI of Principle Investigator);
- Voor nummer zes geldt: Afdelingen en secties hebben hoofden die voor hun afdeling/sectie beleid bepalen als het gaat om onderzoek, meestal zijn dat ook PI's;
- Voor zes en zeven geldt: De data stewards werken voor faculteiten onder aansturing van de bibliotheek. Zij proberen algemeen beleid t.o.v. de behandeling van data te formuleren en verspreiden en adviseren de onderzoekers hierin;
- Voor acht geldt: Met name IT ondersteuners in de afdelingen/secties van faculteiten houden zich bezig met waar welke data opgeslagen zouden moeten worden en adviseren en ondersteunen de onderzoekers hierbij;
- In verschillende secties wordt mogelijk gebruik gemaakt van externe diensten voor de opslag van data. Voorbeelden daarvan worden wellicht tijdens de empirische fase duidelijk.
- Computational stakeholders zijn computers die zelf data verzamelen en interpreteren (op basis van de metadata), hiervan zijn voorafgaand aan de interviews geen voorbeelden bekend op de case organisatie.

Wat vervolgens weer resulteert in de lijst:

- Principle Investigators (PI's);
- Onderzoekers;
- Data Stewards;
- IT ondersteuners (in de afdeling/sectie).

10.2. Beschrijving geïnterviewden

Er zijn vier manieren gebruikt om kandidaten voor interviews te vinden:

- Via een organisatiebreed orgaan waarvan de leden zich bezighouden met High Performance Computing, bij uitstek een vakgebied waarin veel data gebruikt wordt;
- Via de data stewards, zij houden een lijst met zogenaamde 'Data Champions' bij. Dit zijn mensen, meestal onderzoekers, die hebben aangegeven dat ze graag meedenken en meepraten over databeheer;
- Eveneens via de data stewards, maar niet van de Data Champion lijst (twee data stewards en een IT ondersteuner);
- Eén kandidaat die in eerdere projecten ondersteuning bij ICT heeft gevraagd bij de opslag en het beheer van zijn data.

Hieronder staat de lijst met geïnterviewden. Ze staan niet op chronologische volgorde van afgenomen interview Alleen de functietitel en de faculteit wordt gegeven uit privacy overwegingen. Dezelfde volgorde is gebruikt in Atlas en bij het indexeren van de interviews. Er zijn twee imports niet in één keer goed gegaan, daardoor vallen de nummer 4 en 13 uit.

1. Assistant professor, Data Champion, faculteit 1 (via data stewards);
2. Data steward, faculteit 2 (via data stewards);
3. Assistant professor, Data Champion, faculteit 3 (via data stewards);
4. Uitgevallen, zie boven;
5. Assistant professor, Data Champion, faculteit 4 (via data stewards);
6. Data steward, faculteit 1 (via data stewards);
7. Professor, Data Champion faculteit 1 (via data stewards);
8. Twee geïnterviewden: een PHD student en docent en een assistent professor, Faculteit 5 (via HPC);
9. Professor, faculteit 2 (via HPC);
10. Associate professor, faculteit 4 (via HPC);
11. Professor, Data Champion, faculteit 6 (via data stewards);
12. Postdoc researcher, Data Champion, Faculteit 5 (via data stewards);
13. Uitgevallen, zie boven;
14. Assistent professor, faculteit 1 (via HPC);
15. Technical ICT Support Research and Education, faculteit 1 (via data stewards);
16. Professor, faculteit 7 (via ICT).

Dit resulteert in TABEL 34:

Lijst met Geïnterviewden en hun invloed binnen een onderzoek							
#	Rol	Rol in onderzoek	Benaderd via	Data Champion?	Bepaalt RDM Richtlijnen	Adviseert RDM Richtlijnen	Onder-steunt onderzoek
1	UD	PI	DS	Ja	Groep	Groep	Groep
2	DS	/	DS	Nee	Fac/groep	Fac/groep	Fac/groep
3	UD	PI	DS	Ja	groep	groep	groep
5	UD	PI	DS	Ja	groep	groep	groep
6	DS	/	DS	Nee	Fac/groep	Fac/groep	Fac/groep
7	HL	PI	DS	Ja	Afd/groep	Afd/groep	Afd/groep
8	UD	PI	HPC	Nee	groep	groep	groep
9	HL	PI	HPC	Nee	Afd/groep	Afd/groep	Afd/groep
10	UHD	PI	HPC	Nee	groep	groep	groep
11	HL	PI	DS	Ja	Afd/groep	Afd/groep	Afd/groep
12	Postdoc	Onderzoeker	DS	Ja	/	grp	grp
14	UD	PI	HPC	Nee	Fac/groep	Fac/groep	Fac/groep
15	Support	Support	DS	Nee	/	/	Afd/groep
16	HL	PI	ICT	Nee	Afd/groep	Afd/groep	Afd/groep

Opmerkingen: De tabel geeft aan waar personen mogelijk invloed kunnen uitoefenen en hoe. Daar waar PI genoemd staat, betekent dat de persoon als PI kan optreden, dat wil niet zeggen dat dat ook altijd zo is.

Professors, Associate Professors en Assistant Professors doen of deden allen onderzoek en/of geven leiding aan mensen die onderzoek doen. Zij kunnen allemaal als PI optreden. De professor is ook leerstoelhouder of afdelingshoofd en kan in die zin RDM beleid bepalen voor iedereen in de afdeling.

Een PI is altijd ook een onderzoeker. Een PI kan beleid bepalen rondom RDM in de onderzoeksgroep, maar kan er ook voor kiezen te adviseren en/of te ondersteunen. Dat is redelijk vrij.

De professor zal meer aan de bepalende kant staan van de RDM richtlijnen in de onderzoeksgroep/afdeling, een Associate en Assistant Professor meer aan de adviserende en ondersteunende kant.

Afkortingen in de tabel:
HL = hoogleraar => Professor
UHD = Universitair HoofdDocent => Associate Professor
UD Universitair Docent => Assistant Professor
DS = Data Steward
Fac = faculteit
Afd = Afdeling binnen een faculteit
Grp = Onderzoeksgroep

Tabel 34, lijst met geïnterviewden en hun invloed binnen een onderzoek

10.3. Uitnodiging interviews

(Uitnodiging via data stewards in het Engels omdat niet elke onderzoeker van Nederlandse afkomst is).

Dear Colleague,

Please allow me to introduce myself. My name is Bert Kuipers and I work for the ICT department as manager of the Systems department. A few years ago I took it on me to do a masters in Business Process Management and IT. My final thesis is about research data management, specifically the management of research data in the active phase of research.

My problem definition is:

Which design principles should a Research Data Management work system, focused on the management of active research data, within your organization, meet so that these data be findable, traceable and archivable?

(The concept of a work system describes how people use information, technological aids, information systems or any other aids to get certain work done).

Would you be willing to help me by letting me interview you on your usage of RDM in the active phase of your research? My goals are twofold:

- I would like to verify if my findings from the literature review are correct.
- I would like to learn what you are doing differently in your day to day business from what I found in literature.

Examples of questions could be:

- What do you consider active research data?
- What would be reliable data storage? What are your needs?
- Is it important to synchronize data from external sources to a central storage location?
- With whom would you share your data, what are their locations?

I think we would need an hour for the interview. If there are no objections, I would like to make an audio recording of the interview. Audio will be destroyed after the transcription is done.

Thank you and with

Best Regards,

Bert

10.4. Initieel script interviews

Interview

Uit eerdere interviews op de TU zijn de volgende problemen naar voren gekomen:

1. Alle data van een onderzoek (of onderzoeker) staan verspreid over diverse systemen, vaak ook op diverse locaties, zonder dat zichtbaar gemaakt kan worden welke data bij elkaar horen.
2. Men zou graag alle relevante data van een onderzoek op eenvoudige wijze in het 4TU datacenter willen archiveren.

Eerste vragen:

1. Herkent u dit probleem (stelling 1)?
    - Heeft u daar wat aan toe te voegen?
2. Is dit ook uw wens (stelling 2)?
    - Heeft u er wat aan toe te voegen?

Er worden 3 definities gegeven van aspecten waar de research data aan moeten voldoen om bovenstaand probleem op te lossen.

Vindbaarheid (voor dit onderzoek): De bijbehorende vraag is: ”Waar zijn de data (fysiek en/of logisch)?”

Volgbaarheid (voor dit onderzoek): De bijbehorende vraag is: ”Kan ik de wijzigingen die de data ondergaan volgen op basis van de beschikbare informatie?” Het gaat hierbij om de gewijzigde data zelf, de informatie over de wijzigingen en de hulpmiddelen die helpen de wijzigingen te herkennen.

Archiveerbaarheid (voor dit onderzoek): De bijbehorende vraag is: ”Wat zijn de relevante data die gearchiveerd moeten worden en hoe kan ik ze als dataset herkennen?”

Als data op zodanige wijze wordt aangemaakt, bewerkt , opgeslagen en wellicht ook vernietigd dat het aan bovenstaande aspecten blijft voldoen, dan wordt voldaan aan stelling 1 en 2.

3. Bent u het eens met dat uitgangspunt?
    - Heeft u er wat aan toe te voegen?

Wij zijn op zoek naar de specificaties van een werksysteem dat ACTIEVE research data managet en er voor zorgt dat ze vindbaar, volgbaar en archiveerbaar zijn. Voor nu kan het werksysteem als een black box gezien worden waar opdrachten in gestopt worden en resultaten uitkomen.

Voorbeelden van opdrachten en resultaten zijn:

- Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan;
- Het werksysteem keurt de aanvraag en kent ruimte toe of niet;
- De researcher bewerkt (opslaan, wijzigen, verplaatsen, verwijderen) zijn (meta)data op de geboden opslagruimte;
- De researcher bedenkt en gebruikt een heldere mappenstructuur op het werksysteem;
- Het werksysteem synchroniseert lokale onderzoeksdata en metadata met de centraal aangeboden opslagruimte;
- Het werksysteem voert versiebeheer uit over de (meta)data;
- Het werksysteem levert de researcher mogelijkheden voor de aanmaak van metadata;
- De researcher vraagt een PID aan bij het werksysteem voor een relevante datasets en bijbehorende metadata;

4. Vindt u deze set aan activiteiten herkenbaar?
5. Zijn er belangrijke activiteiten die we mogelijk zijn vergeten?

6. Er volgen nu 15 stellingen die globale functionele eisen en wensen aan het werksysteem beschrijven. Er is hierin rekening gehouden met vindbaarheid, volgbaarheid en archiveerbaarheid. Aan u de vraag ze door te lezen en te prioriteren naar vereist, zeer gewenst, gewenst en niet nodig. Let wel het gaat om een werksysteem voor **actieve** data.

    1. Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.
    2. Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders.
    3. Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.
    4. Het werksysteem beschrijft (meta)Data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR principes volgen.
    5. Het werksysteem ondersteunt het aanmaken van metadata (bij voorkeur geautomatiseerd).
    6. Het werksysteem verzekert dat backupmogelijkheden gedurende het onderzoek beschikbaar zijn.
    7. Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.
    8. Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen
    9. Het werksysteem levert de mogelijkheid om data en bijbehorende metadata en/of datadocumentatie te vernietigen
    10. Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.
    11. Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering).
    12. Het werksysteem stelt (meta)Data met een duidelijke en toegankelijke dataovereenkomst beschikbaar.
    13. Het werksysteem kan unieke identifiers aanmaken voor relevante onderzoeksdata (Persistent Identifiers, PID). De wetenschapper bepaalt hierbij wat relevante onderzoeksdata zijn.
    14. Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.
    15. Het werksysteem heeft een userinterface en API om informatie te tonen, activiteiten te volgen, logs te lezen en (zoek)opdrachten te geven.

7. Vind u dat we stellingen missen?
8. Zijn er stellingen waar u op terug zou willen komen?
9. Welke stellingen hebben naar uw mening topprioriteit?
10. Hoe zou een ideaal werksysteem er volgen u uitzien (wie zijn er bij betrokken, welke activiteiten voeren ze uit, welke hulpmiddelen gebruiken ze, welke informatie hebben ze nodig)?

10.5. Uitwerking totstandkoming concepten voor interviews

Het literatuuronderzoek levert 14 ‘products and services’ bevindingen, gesorteerd op vindbaar, volgbaar en archiveerbaar. De concepten en eigenschappen zijn geel gemarkeerd:

Vindbaar

- Het werksysteem levert (toegang tot) een doorzoekbare bron waarin data met de daarbij behorende metadata en/of datadocumentatie vindbaar zijn.
- Het werksysteem biedt een betrouwbare centrale opslagplaats met voldoende capaciteit die gesynchroniseerd kan worden (eenmalig, periodiek of continue) met de opslag van externe onderzoeksdatahouders.
- Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.

Vindbaar en volgbaar

- Het werksysteem beschrijft (meta)data met voldoende relevante attributen en een taal die formeel, toegankelijk, gedeeld en wijds beschikbaar is en bovendien de vocabulaires heeft die de FAIR (Findable, Accessible, Interoperable, Reusable) principes volgen.
- Het werksysteem ondersteunt het aanmaken van metadata (bij voorkeur geautomatiseerd).
- Het werksysteem verzekert dat restoremogelijkheden gedurende het onderzoek beschikbaar zijn.
- Alle bewerkingen met betrekking tot vindbaarheid en volgbaarheid van de data (verplaatsen, kopiëren, vernietigen, wijzigen, aanmaken) worden door het werksysteem gelogd, waarbij de logs doorzoekbaar zijn.

Volgbaar

- Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen.
- Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.

Volgbaar en archiveerbaar

- Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering).

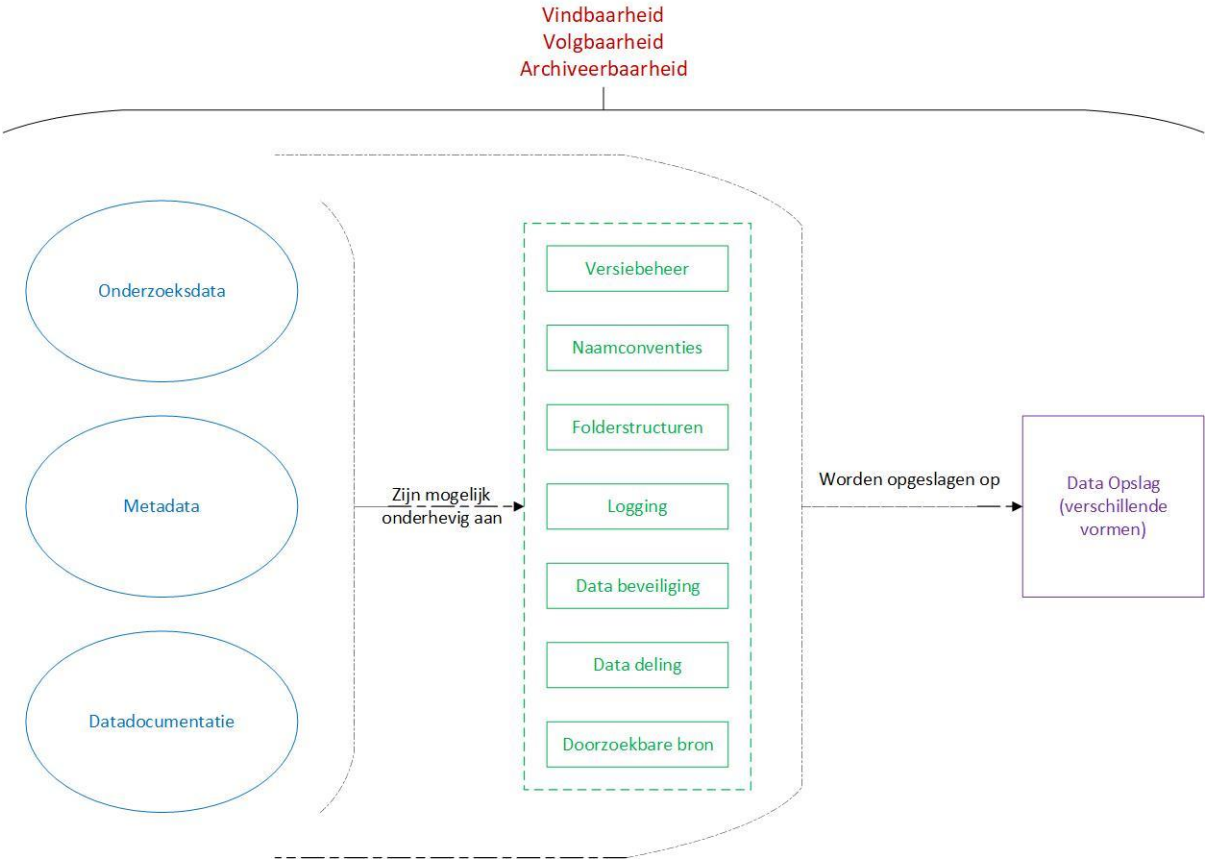
Archiveerbaar

- Het werksysteem stelt (meta)data met een duidelijke en toegankelijke dataovereenkomst beschikbaar.
- Het werksysteem kan unieke identifiers aanmaken voor relevante onderzoeksdata (Persistent Identifiers, PID). De wetenschapper bepaalt hierbij wat relevante onderzoeksdata zijn.

Vindbaar, volgbaar en archiveerbaar

- Het werksysteem kan actieve (meta)data beschikbaar stellen aan de internationale onderzoeksgemeenschap, waarbij de veiligheid en integriteit van de (meta)data gewaarborgd worden.
- Het werksysteem heeft een userinterface en API om informatie te tonen, activiteiten te volgen, logs te lezen en (zoek)opdrachten te geven.

De twee ‘technology’ bevindingen leveren verder geen andere concepten op. Bovenstaande wordt samengevat in een conceptschema (FIGUUR 20) en bijbehorende vragen per conceptgroep:



Figuur 20, conceptualisering theorie

10.6. Uiteindelijk script voor interviews

In principe staat de volgorde niet vast. Het lijkt wel verstandig om te beginnen met het vragen naar onderzoeksdata (zoals: wat zijn het, wanneer zijn ze actief) en daar de rest van het interview aan op te hangen. Als in een antwoord een bepaalde richting wordt gegeven (bijvoorbeeld delen) dan kan naar dat concept gesprongen worden.

Actieve onderzoeksdata, metadata en datadocumentatie.

Welke onderzoeksdata zijn actieve data gedurende een onderzoek? (onderzoeksdata)
Hoe ziet u de verschillen tussen onderzoeksdata, metadata en datadocumentatie? (onderzoeksdata, metadata en datadocumentatie)
Waar staan u data allemaal opgeslagen? (Als het niet wordt genoemd ook vragen naar de datahouders) (zegt iets over de bestaande infrastructuur) (onderzoeksdata, metadata en datadocumentatie)
Hoe en waar slaat u de relevante data op? (zegt iets over de bestaande infrastructuur, dataopslag, datahouders, onderzoeksdata, metadata en datadocumentatie)
Hoe zou in uw ogen optimale data opslag (voor actieve data!) er uit zien, waar zou het aan moeten voldoen? (dataopslag, datahouders, onderzoeksdata, metadata en datadocumentatie)
Wat zijn in uw ogen relevante data? (onderzoeksdata, metadata en datadocumentatie)
Hoe zorgt u er voor dat u uw data, metadata en datadocumentatie kunt vinden? (onderzoeksdata, metadata en datadocumentatie)

Datadeling

Werkt u ook veel samen op dezelfde datasets (inclusief metadata en datadocumentatie)? (delen, onderzoeksdata, metadata en datadocumentatie)
Zo ja:
Op welke manier geeft u de anderen toegang tot de data? (delen, databeveiliging, onderzoeksdata, metadata en datadocumentatie)
Met wie deelt u de data (binnen de TU, in NL, in de EU, overal)? (delen, onderzoeksdata, metadata en datadocumentatie)
Hoe houdt u dan bij wie wat verandert? (databeveiliging, data volgen, onderzoeksdata, metadata en datadocumentatie)

Databeveiliging

In hoeverre is het voor u van belang dat u uw actieve onderzoeksdata kunt afschermen voor anderen? (databeveiliging, onderzoeksdata, metadata en datadocumentatie)
Subvraag: geldt dat in dezelfde mate voor metadata en datadocumentatie? (databeveiliging, onderzoeksdata, metadata en datadocumentatie)
Hoe zorgt u er voor dat de data consistent blijven gedurende het onderzoek? Dat u altijd weet wat er met uw data gebeurt? Dat u weet welke wijzigingen er hebben plaatsgevonden. (databeveiliging, data volgen, onderzoeksdata, metadata en datadocumentatie)
Subvraag: Is het wel eens nodig om de data te transformeren (bijvoorbeeld anonimiseren) voor u ze met de buitenwereld deelt? (data volgen, onderzoeksdata, metadata en datadocumentatie)
Hoe zorgt u er voor dat u weer terugkunt naar een dataset waar u zeker van bent dat hij consistent is? (databeveiliging, data volgen, onderzoeksdata, metadata en datadocumentatie)

Doorzoekbare bron en interface

Hoe zorgt u er voor dat u uw data, metadata en datadocumentatie kunt vinden? (doorzoekbare bron, userinterface, api, vinden)

Hoe zorgt u er voor dat u alle wijzigingen die aan een dataset (of metadata of datadocumentatie) plaatsvinden kunt volgen/terugzoeken? (doorzoekbare bron, userinterface, api, volgen)
Hoe zou vinden en volgen voor u het makkelijkst gaan? (vinden en volgen)
Indien nog niet genoemd:
Maakt u hiervoor gebruik van naamconventies en een mappenstructuur? Hoe ziet dat er uit? Welke voordelen geeft het? Hoe zou dit optimaal zijn? (structuur, naamconventies)
Maakt u gebruik van versies? Hoe doet u dat? Hoe zou dit optimaal zijn? Zou dat in uw ogen geautomatiseerd kunnen? (versies)
Geeft u uw datasets een unieke identifier? Zou het handig zijn als het systeem er 1 voor u kon genereren? (PID)

11. Bijlage uitwerking interviews

In deze bijlage wordt per codecategorie uit de interviews uitgewerkt wat de definities zijn die bij de case organisatie gehanteerd worden en welke relaties codes met elkaar hebben. De bijlage kan als codeboek dienen.

11.1. Onderzoeksdata

Uit de literatuurstudie : 'feitelijke data (zoals numerieke scores, tekstuele records, afbeeldingen en geluiden) die worden gebruikt als primaire bronnen voor wetenschappelijk onderzoek en die algemeen aanvaard zijn in de wetenschappelijke gemeenschap om de onderzoeksresultaten te valideren'.

Onderzoeksdata hebben drie belangrijke eigenschappen:

1. Ze dienen als primaire bron voor wetenschappelijk onderzoek;
2. In de wetenschappelijke gemeenschap worden ze geaccepteerd als noodzakelijk om wetenschappelijke bevindingen te valideren;
4. De onderzoeker bepaalt wat de relevante data van een onderzoek zijn.

In de paragrafen hieronder wordt uitgewerkt wat er in de interviews over research data naar voren kwam. Het begint met de definitie, dan wat actieve data zijn, waar de data vandaan komt, de niveaus (levels), specifieke soorten data als meetdata, simulatiedata en code, research data en lab notebooks, het opruimen van research data en de gevoeligheid van research data. Er wordt afgesloten met een conclusie.

Bij het coderen van de interviews kwamen de volgende relevante codes naar voren (TABEL 35).

Subcodes in de codecategorie Research Data	
Subcode omschrijving	Subcode in Atlas
De definitie van research data	rd definitie
Wat actieve onderzoeksdata zijn	rd actief
Waar de data vandaan komen/uit voortkomen	rd herkomst
Welke verschillende niveaus van onderzoeksdata er zijn	rd levels
Welke soorten onderzoeksdata er zijn	rd meetdata, rd simulatie, rd code, rd eln
Of onderzoeksdata wordt opgeruimd	rd opruimen
Wanneer onderzoeksdata gevoelig zijn	rd gevoelig

Tabel 35, Subcodes in de codecategorie Research Data

11.1.1. Onderzoeksdata definitie

Vanuit het beleid van de case organisatie over wat gearchiveerd moet worden, wat sommige wetenschappelijke tijdschriften nu vragen en de samenvatting van geïnterviewde 16 kan het volgende gesteld worden over onderzoeksdata in het kader van RDM:

Onderzoeksdata voor RDM in de actieve fase zijn alle data die nodig zijn om het onderzoek de eerste keer uit te kunnen voeren en later te kunnen herhalen (2:78, 9:6 en 16:27[23]). De afzonderlijke onderdelen van de definitie staan in TABEL 36.

[23] 2:78 Alle onderzoeksdata, code en documentatie die nodig zijn om het onderzoek te herhalen, moeten gearchiveerd worden in het 4TU datacenter (TU beleid).
9:6 Alle data die benodigd zijn om het onderzoek te herhalen moeten ingeleverd worden bij het tijdschrift (om het onderzoek te kunnen beoordelen en om het te kunnen herhalen).

Definitie onderzoeksdata en opdeling in soorten onderzoeksdata, rd definitie			
	Genoemd in	#gen	#int
Definitie: Onderzoeksdata in het kader van RDM zijn alle data die nodig zijn om het onderzoek de eerste keer uit te kunnen voeren en later te kunnen herhalen.	2:78, 9:6 en 16:27	3	3
Feitelijke data zoals benoemd in de definitie uit de literatuur	1:5 3:10, 5:2, 5:3, 5:12, 6:2, 6:3, 6:4, 8:1, 8:2, 10:5, 16:69	12	7
Procedures en analyses uit lablogboeken (voor zover aanwezig)	6:2, 6:3, 6:4	3	1
Scripts/code/software die worden gebruikt om data te creëren of te bewerken	11:1, 11:2, 11:5, 11:12, 14:2, 14:3, 9:8, 15:26, 15:27 16:12, 16:6, 16:8, 16:9	13	5
De data voor de tabellen en figuren in de publicaties	1:5, 3:10, 3:67, 7:9, 10:9, 11:8, 12:13, 16:2	8	7
De publicaties zelf	11:18	1	1
Voorbeelden van quotes: 5:12: Geïnterviewde: dus voor mij is onderzoeksdata vooral vragenlijsten, case studies, interviews. 6:2: Interviewer: wat is voor jou onderzoeksdata? Geïnterviewde: Alle omschrijvingen van sample materiaal dat zijn de questionnaires of surveys die mensen in hebben gevuld, wel anoniem over het grootste gedeelte. Maar ja daar komt ook informatie en data uit rollen. Dat zijn alle notities die ik heb gemaakt over alle procedures in het logboek, in het lab dus, en het algemene lab boek de procedures et cetera die ik moet volgen. Alle fysieke samples eigenlijk ook dat valt er ook onder. Dus dat kunnen dan tanden zelf zijn in buisjes met labels tot op opgeloste tanden in het lab ook in potjes om het zo maar te noemen. Met daar een code op (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS**ERROR! NOT A VALID RESULT FOR TABLE.**).			

Tabel 36, Definitie onderzoeksdata en opdeling in soorten onderzoeksdata

11.1.2. Onderzoeksdata levels

Aansluitend op het voorgaande spreken 10 van de 14 geïnterviewden in enige mate van verschillende niveaus van onderzoeksdata. Van de meest onbewerkte vorm, tot 'het tabelletje in de publicatie'. Er worden drie niveaus voor data onderscheiden (TABEL 37).

In TABEL 36 is voor de publicatie zelf weinig steun. Daarentegen zijn de data die nodig zijn om tabellen en figuren in de publicatie te vullen, volgens beide tabellen van belang. Het lijkt derhalve gerechtvaardigd om de publicatie te lezen als de data die nodig zijn voor de publicatie. De publicatie verdwijnt als onderzoeksdatasoort. Slechts één geïnterviewde noemt procedures en analyses uit lablogboeken. Het zou kunnen dat deze geïnterviewde dit voor heel specifiek onderzoek bedoelt. Als een computerscript wordt gebruikt om data te genereren en te analyseren en als ook computerscripts worden gezien als onderzoeksdatasoort, dan lijkt het redelijk om ook procedures en analyses uit lablogboeken als een onderzoeksdatasoort te zien.

16:27 Uiteindelijk zijn het allemaal bytes (data, software, publicaties, plaatjes en tekst in publicaties). Als je het zo beschouwt, is data management breder.

Onderzoeksdata levels, rd levels			
	Genoemd in	#gen	#int
De **primaire** onderzoeksdata[24] zijn de data die je direct na het ontstaan hebt verkregen of de runtime data die je gedurende een periode tijdens het onderzoek verzamelt	1:5, 3:1, 3:35, 3:67, 7:10, 14:6, 15:8	7	5
De **intermediate** onderzoeksdata ontstaan na bewerking van (analyse op) de primaire data, of na bewerking van eerder verkregen intermediate data, het kunnen ook tussentijdse snapshots van in de tijd veranderende data zijn	3:1, 7:29, 7:34, 7:35, 7:36, 7:37, 7:38, 7:39, 9:2; 11:4, 11:8, 11:16, 14:9,16:7	14	6
De **final** onderzoeksdata zijn de data die nodig zijn om de publicatie te schrijven (data voor figuren, tabellen e.d.)	1:5, 3:10, 3:67, 7:9, 10:9, 11:8, 12:13, 16:2	8	7
Voorbeelden van quotes: 1:5: Geïnterviewde: So that when you say they give the data set that can have several levels? Yeah because if you make a figure that is data right? That's like that level minus what if you say here is a figure. But here is also the actual numerical data that is plotted on the figure. That's an extra level of the data. If you say I'm also providing the raw data and the scripts that I use that's again a difference…… 3:1: Geïnterviewde: Biologische data meten we aan de hand van high throughput instrumenten waardoor we gigantische bergen data hebben en die moeten verwerkt worden door analyses. En dan krijg je intermediate data waarvan je moet bijhouden wat je ermee wil doen. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 37, Onderzoeksdata levels

11.1.3. Onderzoeksdata actief

In eerste instantie werden met actieve data, de data die gedurende het onderzoek gegenereerd en gebruikt worden, bedoeld. Dat is echter geen afdoende definitie, zo blijkt het uit het commentaar van geïnterviewde 3 en 15. Zij stellen dat data actief zijn zolang ze nog door iemand gebruikt worden. Als data is gearchiveerd maar ze worden nog steeds voor (ander) onderzoek gebruikt, dan is het actief.

- **Data zijn actief zolang ze nog door iemand gebruikt worden. Als data zijn gearchiveerd maar nog steeds voor (ander) onderzoek worden gebruikt, dan zijn ze actief** 3:15, 3:16, 3:17, 15:1, 15:2 (vijf keer benoemd in twee interviews).

[24] T.a.v. meetdata: 7:34, 7:35, 7:36, 7:37, 7:38, 7:39 (enigszins: 11:9, 16:14) Wat is precies ruwe data (primaire data)? Een meetinstrument doet vaak al een eerste bewerking van de gemeten waarden. Dat is acceptabel, omdat die bewerking bekend is. Als een volgende bewerking ook bekend is, dan is de vraag waar je de grens moet leggen van wat je primaire data noemt. Stelling als je het pad van de ruwste vorm van data naar de uiteindelijke set precies kunt beschrijven, dan is die uiteindelijke dataset voldoende om op te slaan, inclusief de beschrijving. Dit maakt het onderscheid tussen primaire en intermediate data wel diffuus.

11.1.4. Onderzoeksdata herkomst

Data herkomst is op te splitsen in waar de data letterlijk vandaan komen en hoe de data zijn ontstaan. Het staat samengevat in TABEL 38.

Onderzoeksdata herkomst, rd herkomst			
Herkomst	Genoemd in	#gen	#int
Binnen de TU	3:6, 3:7	2	1
hergebruik van eigen data	15:2	1	1
Vanuit publieke bronnen	3:3, 10:5, 10:6, 16:6, 16:11, 16:48	6	3
Partners in het onderzoek	3:4, 10:5, 10:6, 14:7, 15:3	5	4
Bedrijven	15:3	1	1
Universiteiten	3:6, 3:7, 15:3	3	2
Ziekenhuizen	1:11, 14:4, 14:7, 15:3	4	3
Voorbeelden van quotes: 14:7: Geïnterviewde: CT natuurlijk ook, die verzinnen we ook niet zelf, meestal komt die ergens vandaan. Komt meestal van samenwerking met Erasmus of hier met de kliniek. 15:3: Geïnterviewde: De data komt van bedrijven, universiteiten en ziekenhuizen waar we mee samenwerken en daar zijn ook allemaal verschillende regels omheen. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 38, Onderzoeksdata herkomst

De tweede rij spreekt van eigen data. Eigen data zijn data die van de universiteit zijn. Dat betekent dat de eerste twee bullets kunnen samengevat worden in Binnen de TU.

Buiten de TU kunnen data uit publieke bronnen komen, maar eventueel ook bij partners in het onderzoek (dit kunnen ziekenhuizen, bedrijven en universiteiten zijn). Data kan ook bij andere universiteiten en ziekenhuizen vandaan komen, zonder dat ze partner in het onderzoek zijn.

Ofwel onderzoeksdata komen van binnen en buiten de case organisatie. Data van buiten de case organisatie kan van academische en niet academische instellingen komen.

Een andere vorm van herkomst is hoe de data zijn ontstaan (uit metingen, simulaties, code, ELN) dit staat beschreven in de volgende tabellen (TABEL 39, TABEL 40, TABEL 41 en TABEL 42).

Meetdata als onderzoeksdata, rd meetdata			
	Genoemd in	#gen	#int
Definitie: Meetdata zijn data die verkregen worden met meetinstrumenten.	1:11, 2:47, 2:56, 2:81, 2:82, 3:1, 3:2, 6:30, 7:63, 14:34, 14:7, 15:6, 15:9	13	7
Meetdata kunnen realtime zijn in een onderzoek	8:2, 8:3, 15:10	3	2
Meetdata kunnen uit gebruikerstesten voortkomen	16:4	1	1
Meetdata kunnen uit performancetesten voortkomen	16:4	1	1
De data kunnen uit grote en kleine bestanden bestaan en het kan om grote en kleine hoeveelheden gaan	15:6, 15:7	2	1
Voorbeelden van quotes: 8:2: So for me the experimental side of what I was doing here, I was performing a test on a steel bridge, where a vehicle would pass and I would collect information per second. So that is like a huge database. Within three months of terabytes of data and I had one hundred and six sensors installed, which was giving me per second data, different types of data, one was giving me longitudal information one in the vertical direction and one was sheer direction. So that's why we had lots of different information and the numbering and the things that we were getting were very very crucial. Sometimes the device would just stop because the pressure goes off or something happens. But then you need to really filter out the information, otherwise you see suddenly there is a huge strain, so huge that the information doesn't make sense. And then you go back and you check the vehicle just to stop there. So that's why I think this is very important. 16:4: Geïnterviewde: Misschien met gebruikers bruikbaarheid, usability, misschien op performance getest, daar worden meetdata verzameld. En al die andere proeven daarvoor, ja dat zijn dan deelproeven afgebroken proeven waar wel wat metingen zijn die niet oninteressant zijn juist die alternatieven zijn wel aardig om achteraf te beschouwen. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 39, Meetdata als onderzoeksdata

De derde en vierde rij zeggen vooral wat over de herkomst van meetdata en worden voldoende gedekt bij de conclusies over herkomst van data. Dat gemeten data realtime kan zijn, zou van invloed kunnen zijn op later te maken ontwerpkeuzes (bijvoorbeeld de wegschrijfsnelheid van de opslagsystemen). Dat geldt ook voor het kunnen omgaan met verschillende bestandsformaten en verschillende hoeveelheden.

Meetdata zijn data die verkregen worden met meetinstrumenten. Het komt voor dat ze realtime worden vergaard. Hoeveelheden data en bestandsformaten kunnen sterk wisselen.

Gesimuleerde data als onderzoeksdata, rd simulatie			
Soorten simulatiedata	Genoemd in	#gen	#int
Simulatiedata zijn data die gegenereerd worden met de computer	1:9, 1:15, 1:16, 9:1, 9:3 11:3, 11:11, 14:5, 15:9, 15:11	10	5
Data die worden gebruikt voor onderzoek kunnen modellen van werkelijke omstandigheden zijn, dit zijn gesimuleerde data	1:15, 1:16	2	1
Voorbeelden van quotes: 1:15: Geïnterviewde: So you can also use some what we call kind of toy problems and you kind of make up models. So you would make up basically a little anatomy where like your spinal cord would be just a cylinder and then a tumor around would be perfect the half cylinder things like that so that then you kind of make up a model yourself. 9:1: Interviewer: Hoe komen jullie aan je data. Komt dat uit metingen of simulaties? Geïnterviewde: We genereren met onze simulaties behoorlijk wat data. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 40, Gesimuleerde data als onderzoeksdata

De tweede regel uit TABEL 40, is een verbijzondering van de eerste. Het wordt slechts in één interview zo benoemd. Voor simulatiedata wordt daarom volstaan met de beschrijving op de eerste regel: Simulatiedata zijn data die gegenereerd worden met de computer.

Code als onderzoeksdata, rd code			
Aspect van code data	Genoemd in	#gen	#int
Code is een datacategorie die van belang is voor het onderzoek	1:5, 1:10, 3:10, 15:26, 15:27, 16:7, 16:9	7	4
De code is voor software/scripts die de modellen beschrijven die worden gebruikt voor het onderzoek	11:5, 11:12, 14:2, 14:8	4	2
Voor sommige onderzoeken is de code belangrijker dan de data zelf (met de code kan data opnieuw gegenereerd worden)	9:3, 9:6, 11:5	3	2
Code data kunnen noodzakelijk zijn om het onderzoek te herhalen	9:6, 9:8, 10:9, 11:2	4	3
Voorbeelden van quotes: 11:5: Interviewer: So the scripts to create the data are part of the data as well. Geïnterviewde: I would say so yes, because the data only makes sense, if you know how it was created. So the interpretation, otherwise it's just numbers. 9:8: Interviewer: En die scripts en de parameters die je erin stopt zou je dat onderzoeks data noemen? Of zeg je dat vind ik iets wat ik data documentatie zou noemen? Geïnterviewde: Ja, kwestie van definitie denk ik, ik zou het toch wel onderzoeksdata willen noemen want je hebt het gebruikt bij je onderzoek, dus wat we wel publiceren is alles wat je theoretisch nodig zou moeten hebben om het onderzoek te kunnen herhalen, maar niet bijvoorbeeld echt alle individuele scripts. Dus mensen zouden die dus zelf kunnen schrijven met de informatie uit het artikel. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 41, Code als onderzoeksdata

De inhoud van TABEL 41 kan als volgt samengevat worden: Code data kunnen bestaan uit de code voor scripts en programma's die worden gebruikt om (simulatie)data te genereren en/of modellen beschrijven die worden gebruikt in het onderzoek. De code kan belang zijn voor herhaling van (stappen in) het onderzoek.

ELN inhoud als onderzoeksdata, rd ELN			
De data in het lablogboek beschrijven	Genoemd in	#gen	#int
De gebruikte protocollen, het onderzoek(aspect) en de uitkomsten	2:49, 2:51, 12:33, 6:2, 6:3, 6:4	6	2
Waar data staan	3:39	1	1
Hoe data gedurende het onderzoek transformeren	3:37, 3:39, 6:2, 6:3, 6:4	5	2
De omstandigheden rondom het onderzoek	12:2	1	1
Voorbeelden van quotes: 3:39: Geïnterviewde: Dus wat wordt beschreven in het lab notebook is vandaag heb ik deze analyse gedaan en dan kijk je in die directory en dan zie je van die analyse al die log files en scripts en wat nog staan daaruit kun je reconstrueren wat is nu precies gebeurd. 12:2: Geïnterviewde: Temperature and pressure and and the flow rate. But this is where the lab notebook becomes important because you need to keep a note of exactly which fluid rate you're setting. And then you measure the pressure coming out. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 42, ELN inhoud als onderzoeksdata

Regel twee heeft zwakke steun. Het kan echter van groot belang zijn voor vindbaarheid en wordt daarom gehandhaafd. Ook regel vier wordt matig gesteund. Het lijkt van belang om iets te kunnen zeggen over de waarde van de onderzoeksdata en valt daarmee meer eerder onder metadata.

De inhoud van TABEL 42 kan als volgt samengevat worden: Het ELN bevat informatie over de gebruikte protocollen en de wijze waarop de data tijdens het onderzoek transformeren. Ook kan het ELN informatie bevatten over waar de data staan.

11.1.5. Onderzoeksdata Opruimen

Het bestaan van de data eindigt als het wordt opgeruimd, redenen daarvoor staan in TABEL 43.

Onderzoeksdata opruimen, rd opruimen			
Redenen om op te ruimen	Genoemd in	#gen	#int
De data hebben een verloopdatum	1:27, 1:30	2	1
De data zijn eenvoudig opnieuw te genereren	9:32	1	1
Er is ruimte nodig voor nieuwe data	6:59, 6:60, 9:32, 9:33, 15:12, 15:13	6	3
Ze gaan nooit meer gebruikt worden	3:15, 6:59, 6:60, 15:12, 15:13	5	3
Voorbeelden van quotes: 9:32: Interviewer: Goed eigenlijk. Loop je wel eens tegen capaciteitsprobleem aan? Geïnterviewde: Nee, valt wel mee omdat ik gewoon hele grote.... De echte hele grote files die data die datasets die allemaal bewerkt zijn en eh als ze...Als dat klaar is gooien ze weg en dan werken we gewoon met de gereviseerde dataset. Plus het feit dat de code en alle parameters om over te doen.... we een grote dataset altijd opnieuw kunnen genereren. Ik denk dat ik in de 22 jaar dat ik met onderzoek bezig ben, dat ik eh... ja de de echte belangrijke data dat is minder dan een terabyte. Dat is iets wat je gewoon op een normale Mac kan opslaan. 15:13: Geïnterviewde: Daar moet je overigens wel, daar moet je actief actie op nemen. Je kan dat niet overlaten aan de zelfredzaamheid van een onderzoeker, dat moet je sturen, daar moet je opdracht toe geven. Maar dat gebeurt dan, terabytes worden weggegooid, 50, 100 terabyte wordt zo weggegooid. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 43, Onderzoeksdata opruimen

De eerste regel wordt matig ondersteund. Als data een verloopdatum hebben, gaan ze niet meer gebruikt worden. Daarmee wordt de eerste regel afgedekt met de vierde. De tweede regel wordt ook matig ondersteunt. De reden lijkt wel erg logisch en de verwachting is dat andere onderzoekers zich net zo opstellen in soortgelijke situaties.

In het algemeen worden data zo min mogelijk weggegooid (zeven keer benoemd in vijf interviews: 1:39, 2:54, 2:73, 3:36, 6:58, 15:12, 15:14). Als ze worden weggegooid is dat omdat ze niet meer

gebruikt gaan worden. Het kan ook een pragmatische reden hebben: het is eenvoudig nieuwe data te genereren of er is ruimte nodig voor nieuw onderzoek.

11.1.6. Onderzoeksdata gevoelig

Data kunnen om drie redenen gevoelig zijn (TABEL 44).

Onderzoeksdata gevoelig, rd gevoelig			
Onderzoeksdata zijn gevoelig om	Genoemd in	#gen	#int
Medische redenen	3:2	1	1
Privacy redenen	3:2, 5:13, 15:4, 16:69	4	4
Contractuele redenen	10:5, 14:34, 14:35, 15:4	4	3
Voorbeelden van quotes: 3:2: Geïnterviewde: Die originele data kan ofwel van eigen meetinstrumenten komen. In dat geval zijn die bijzonder waardevol, kunnen die gevoelig zijn, kunnen die privacy gevoelig zijn, medisch gevoelig zijn. Je noemt het maar. 15:4: Geïnterviewde: Voor bedrijven geldt vaak een secracy agreement omdat dat concurrentie gevoelig is. Ziekenhuizen werken vaak met privacygevoelige data dus dat is een andere reden waarom je dan voorzichtig moet zijn ermee. En universiteiten dat is vaak wat makkelijker. Dat is vaak wat openbaarder (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 44, Onderzoeksdata gevoelig

Dat onderzoeksdata gevoelig zijn om medische redenen wordt matig ondersteund. Maar buiten geïnterviewde 3 werken ook geïnterviewden 1 en 14 met medische data. Het ligt voor de hand dat ook zij tegen de gevoeligheid van medische data kunnen aanlopen.

11.1.7. Code co-occurrence binnen de onderzoeksdata categorie

De sterkste verwantschappen binnen de research data categorie staan beschreven in (TABEL 45). Waarden tussen de 0.2 en 0.3 worden gezien als mogelijke verwantschappen (eronder betekent zeer waarschijnlijk geen verwantschap en erboven zeer waarschijnlijk wel verwantschap).

Verwantschappen binnen de RD categrorie					
	Code 1	Code 2	Coëfficiënt	#gen	#int
1	rd definitie	rd code	0.28, 7		
2	rd herkomst	rd meetdata	0.27, 9		
Uitleg: 1. De computational scientists (1, 8, 9, 10, 11, 14, mensen die High Performance Computing clusters gebruiken) vinden code van primair belang voor hun onderzoekswerk. Ze noemen het regelmatig tijdens de interviews. Code wordt beschouwd als onderzoeksdata. Deze verwantschap is goed verklaarbaar en geeft geen extra informatie aan de beschrijving van onderzoeksdata. **2.** Veel geïnterviewden geven aan meetdata te gebruiken. Veel data komt van meetinstrumenten. Deze verwantschap is goed verklaarbaar en geeft geen extra informatie aan de beschrijving van onderzoeksdata. De sterkte van een verwantschap (co-occurrence) tussen twee codes wordt uitgelegd in Bijlage Uitleg Co-Occurrence/verwantschap.					

Tabel 45, sterkste verwantschappen binnen de RD categorie

11.2. Onderzoeksdata conclusie

De definitie van research data uit de literatuurstudie wordt bevestigd. Wel constateert men drie verschillende niveaus van onderzoeksdata:

1. Primair: De data zoals die direct na ontstaan zijn verkregen;

2. Intermediate: De data die bewerkt zijn en waar analyses op verricht worden;
3. Final: De data die gebruikt worden om de tabellen en figuren in de publicatie te vullen.

Onderzoeksdata voor RDM in de actieve fase zijn alle data die nodig zijn om het onderzoek de eerste keer uit te kunnen voeren en later te kunnen herhalen, of om delen te kunnen herhalen.

Data komen van binnen en buiten de case organisatie. Data van buiten de case organisatie kan van academische en niet academische instellingen komen.
Data komen voort uit simulaties, code, metingen, bewerkingen van data, uit logboeken en uit gebruikerstesten. Meetdata kunnen realtime worden vergaard. Hoeveelheden data en bestandsformaten kunnen sterk wisselen.

Het bestaan van de data eindigt als het wordt opgeruimd. In het algemeen worden data zo min mogelijk weggegooid. Als ze worden weggegooid is dat omdat ze niet meer gebruikt gaan worden. Het kan ook een pragmatische reden hebben: het is eenvoudig nieuwe data te genereren of er is ruimte nodig voor nieuw onderzoek.

Data kunnen om drie redenen gevoelig zijn: medische, privacy en contractuele redenen.

11.3. Metadata

Volgens de definitie uit de literatuur hebben metadata drie eigenschappen. De eerste is de daadwerkelijke definitie, de tweede beschrijft de vorm en de derde de inhoud.

Metadata:

1. Definitie: Metadata zijn data die informatie geven over de onderzoeksdata met als doel de onderzoeksdata voor anderen bruikbaar te maken (reproduceerbaar en interpreteerbaar) (National Science Board and National Science Foundation 2005).
2. Vorm: Metadata kunnen in lagen geformuleerd worden. De eerste laag is een breed toepasbare standaard (zoals de Dublin Core standaard voor metadata (DCMI 2018)). Deze wordt gevolgd door een domein/discipline specifieke laag en vervolledigd met een laag die de data op folder-, file- en naamniveau beschrijft (Wilkinson et al. 2016).
3. Inhoud: Metadata zouden tenminste moeten bestaan uit een dataset identifier, titel, beschrijving, de maker, contact (persoon/instituut), publicatiedatum, versie, identifiers van de makers en de licentie (Starr et al. 2015).

Bij het coderen van de interviews kwamen de volgende relevante codes naar voren (TABEL 46):

Subcodes in de codecategorie metadata	
Subcode omschrijving	Subcode in Atlas
De definitie van metadata	md definitie
De vorm van metadata	md vorm en md standaard
De inhoud van metadata	
Folderstructuur persoonlijk en groep	md fdst persoonlijk, md fdst groep
Naamconventie persoonlijk en groep	md nmc persoonlijk, md nmc groep
Metadata inhoud (belangrijke velden)	md inhoud
Of men überhaupt metadata gebruikt	md ja

Tabel 46, Subcodes in de codecategorie metadata

11.3.1. Het gebruik van Metadata

De meeste geïnterviewden gebruiken op één of andere manier metadata (TABEL 47).

Metadata gebruiken, metadata ja			
Md ja	Genoemd in	#gen	#int
Alle geïnterviewden geven aan metadata aan te maken in de actieve fase van het onderzoek, behalve geïnterviewde 2	1:17, 1:18, 1:19, 1:20, 1:51, 3:26, 3:27, 3:28, 5:18, 5:22, 5:29, 6:6, 6:7, 7:44, 8:3, 8:16, 8:17, 9:18, 9:19, 9:20, 11:26, 11:41, 11:42, 12:4, 12:5, 15:15, 15:16, 15:17, 15:18, 15:19, 15:20, 16:45	32	11
Voorbeelden van quotes: (3:26 en 3:27): Interviewer: Metadata in de fase voor archivering, laat ik het zo maar noemen, maak je daar gebruik van? Geïnterviewde: Ja. Interviewer: Waar moet ik aan denken wat voor soort.... Geïnterviewde: Een Excel tabel met een unieke identifier met alle kolommetjes met alle metadata die je zou willen. 5:18: Interviewer: Metadata, Maak je daar gebruik van? Geïnterviewde: Ik denk dat het inherent is aan het verzamelen van data, dat je ook metadata hebt. Voor mij klinkt het als: Ja natuurlijk. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 47, Metadata gebruiken

Geïnterviewde 2 (een data steward) geeft aan dat de meeste mensen niet lijken te weten wat metadata is en dat er geen goede voorzieningen zijn om metadata te verzamelen (2:22, 2:24).

11.3.2. Metadata Definitie

Hoewel de meeste geïnterviewden aangeven metadata te gebruiken, is men vager over de definitie en zijn daar ook duidelijk minder quotes van (TABEL 48)

Mtadata definitie, md definitie			
Metadata definitie en inhoud	Genoemd in	#gen	#int
Metadata zijn data die de onderzoeksdata leesbaar, interpreteerbaar en bruikbaar te maken voor anderen	1:19, 3:46, 3:69, 5:20, 5:29, 8:3, 8:17, 9:19, 11:26	9	6
Uit de interviews komt verder naar voren dat de metadata in sommige gevallen ook gegevens bevatten over			
Het functioneren van de meetinstrumenten	8:3, 8:17	2	1
De omstandigheden waarin de data verzameld worden	6:9, 6:10, 15:18, 8:3, 8:16, 8:17	6	3
Voorbeelden van quotes: 8:17: Geïnterviewde 2: And this is crucial for me and that's what I try to inform you about. If you have this kind of information, I see a very weird data here for twenty three point five, and then, because my machine is running separately, and you need to know what happened here, did the machine break or is it real value that you are getting. So you need this one and this one should be coherent. 6:9: Interviewer: Maar resultaten zijn toch geen metadata? Geïnterviewde: In dit geval deels. Omdat de meting van een standaard geeft informatie weer over de andere metingen. Dus als die standaard niet op orde is, dan zegt dat iets over de kwaliteit van de rest van je data. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 48, Metadata definitie

Het eerste deel van de definitie wordt hiermee:
Metadata zijn data die de onderzoeksdata leesbaar, interpreteerbaar en bruikbaar te maken voor anderen. Bovendien kunnen metadata een indicatie geven van de waarde van de verzamelde gegevens en daarmee mogelijk een juiste herhaling van het experiment bevorderen (validiteit).

11.3.3. Metadata vorm

De vorm waarin Metadata beschreven kunnen worden, bestaat volgens de literatuur uit drie lagen (TABEL 49):

Metadata Vorm, md vorm en md standaard			
Metadata in drie lagen	Genoemd in	#gen	#int
De eerste laag is een breed toepasbare standaard (zoals de Dublin Core standaard voor metadata (DCMI 2018))	2:23, 16:45, 16:46	3	2
De tweede laag is een domein/discipline specifieke laag, dit kan een standaard zijn of een gebruik binnen het vakgebied	1:17, 1:51, 2:26, 6:6, 6:31, 16:46, 16:62	7	4
De derde laag beschrijft de data op folder-, naam- en fileformatniveau.	Wordt apart besproken in eigen paragrafen.	/	/
Voorbeelden van quotes: 1:7: Interviewer: OK. And talking about metadata, do you already use metadata in the active phase of your research for your data? Geïnterviewde: I guess you could say try to, so yeah it's some of it is also standardized. So for example for the medical images there is a data standard there's the DICOM standard there. It's a standardized format 16:46: Interviewer: ….. Daarbovenop in onze wereld is het OGC, Open Geospatial Consortium belangrijk die heeft standaarden gemaakt voor de formaten maar ook voor de webservices. Webmapservice, webfeatureservice, van data ophalen. Interviewer: Van metadata? Geïnterviewde: Die hebben ook metadata services, catalog services. Ik zou bijna zeggen dat dat die faciliteiten betreft, metadata, dat de geoinformatie wereld wel 10 20 jaar vooruitlopen op de rest ten aanzien van de informatie voorziening. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 49, Metadata Vorm

Opmerkingen bij TABEL 49: Over de eerste twee lagen kan het volgende geconcludeerd worden:

1. In de interviews worden alle lagen genoemd, maar er is geen één expliciete uitspraak van iemand dat hij/zij alle drie de lagen toepast.
2. De eerste laag (met bijvoorbeeld Dublin Core) wordt door enkele geïnterviewden gebruikt.
3. De (tweede) domein/discipline specifieke laag wordt door enkele geïnterviewden gebruikt.
    a. Er wordt echter ook opgemerkt dat er voor de technische wetenschap relatief weinig standaarden zijn. (2:23)

Ook de derde laag wordt in diverse interviews genoemd. Daar wordt veel meer over gezegd en er blijkt niet 1 werkwijze voor te zijn:

- folderstructuren (

- FOLDERSTRUCTUUR persoonlijk en FOLDERSTRUCTUUR GROEP);
- naamconventies (NAAMCONVENTIE PERSOONLIJK en NAAMCONVENTIE GROEP).

11.3.4. Metadata folderstructuur

Voor het gebruik van een folderstructuur blijken er twee mogelijkheden te zijn:

- een persoonlijke en
- een groepswerkwijze.

Er bestaan ook geen algemene regels of beleid voor. Eén van de datastewards zegt hierover: Naamconventies en folderstructuren zijn niet verplicht in het datamanagementplan (2:46).

Folderstructuur persoonlijk

Folderstructuur persoonlijk, md fdst persoonlijk			
Redenen om een persoonlijke folderstructuur te gebruiken	Genoemd in	#gen	#int
Men heeft de flexibiliteit nodig om per onderzoek een andere structuur te gebruiken	3:60, 9:13, 9:14, 9:29	4	2
Om een structuur te hebben waarin ze alle onderzoeksdata van onderzoeken uit het verleden die ze van belang vinden te bewaren	9:21, 9:22	2	1
Het zo gegroeid is en ze het altijd zo gedaan hebben	1:35, 1:52, 2:63, 5:58, 16:56	5	4
Voorbeelden van quotes: 9:14: Interviewer: Iedereen bepaalt dus zelf hoe die dat eh..... Geïnterviewde: Ja voor elk onderzoek en voor elk type artikel is dat weer anders er is geen uniek formaat. 16:56: Geïnterviewde: En naamconventies, folderstructuren? Daar heb je het net al eventjes over gehad, dat klonk heel persoonlijk vooral. Geïnterviewde: Ja, ad hoc gegroeid. Dus we hebben die M: schijf waar het aardig systematisch opgezet is met dingen met brieven namens de sectie, projecten waar meer mensen aan werken, binnen de sectie. Folder projectnaam en dan zie je een aantal projecten onder die folder. Education folder waar een aantal vakken onder vallen, een folder correspondentie waar de brieven onder vallen een folder notulen sectie overleg, waar de notulen onder vallen. Maar dat is niet geformaliseerd. Het is nu eenmaal zo. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 50, Folderstructuur persoonlijk

Tabel 50 laat zien dat zes geïnterviewden om verschillende redenen een persoonlijke folderstructuur gebruiken.

Folderstructuur groep

Folderstructuur groep, md fdst groep			
Folderstructuren in de onderzoeksgroep	Genoemd in	#gen	#int
Het hebben van een geadviseerde of opgelegde folderstructuur, gebeurt vooral op het niveau van de onderzoeksgroep en veel minder op sectie of afdelingsniveau.	2:41, 2:44, 15:38, 15:39, 16:37	5	3
Een geadviseerde structuur per groep	7:68, 7:71	2	1
Opgelegde structuur per groep	1:35, 1:52, 8:7, 8:9, 8:10, 8:11, 8:31, 10:18	8	3
Voorbeelden van quotes: 7:68: Geïnterviewde: Omdat iedereen er [folders] bij kan. Maar ik laat ze dan in het algemeen best vrij en iedereen heeft zijn eigen manier om dit te gaan organiseren en dat vind ik ook prima 8:11: Interviewer: OK so master students that get here, will have to follow the same structure. Geïnterviewde 3: No we do it for them. So I ask them to give me their data and then I will plug it back in the folder. They don't have access to those confidential to replace (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 51, Folderstructuur groep

Tabel 51 laat zien dat vier geïnterviewden aangeven een soort van groepsfolderstructuur te gebruiken (opgelegd en geadviseerd).

Geïnterviewde 1 geeft in zijn antwoorden aan een eigen (zo gegroeide) folderstructuur te gebruiken die hij ook oplegt aan zijn PHD's en studenten, vandaar dat deze in beide tabellen genoemd wordt.

11.3.5. Metadata naamconventie

Naamconventie persoonlijk

Naamconventie persoonlijk, md nmc persoonlijk			
Verschillende vormen van persoonlijke naamconventies	Genoemd in	#gen	#int
Een volledige persoonlijke invulling	1:49, 2:64, 5:58, 9:13, 9:14, 9:29, 10:14, 12:34, 14:17, 16:56, 16:59	11	8
Een advies van een hoofdonderzoeker waar van mag worden afgeweken	7:65, 7:68, 10:14, 15:38	4	3
Voorbeelden van quotes: 2:64: Geïnterviewde: So then I started to do a year month date and the name of the experiments and I tried to include something in the file name about what it includes so then I know, that was it 10:14: Geïnterviewde: En heb je een eigen naamgeving? Of is dat iets wat jullie met de afdeling gebruiken? Geïnterviewde: Ik heb mijn eigen conventie maar dat is dan bijvoorbeeld ok: n experiments x strategies zoiets weet je wel, dat soort dingen. Interviewer: En je vraagt niet aan je phd's en je master studenten om dat ook te doen? Geïnterviewde: Ik geef ze dat als best practice mee. In die zin wel. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 52, Naamconventie persoonlijk

Invulling is vaak gericht op een duidelijke omschrijving in de bestandsnaam en een datum (ofwel in de bestandsnaam, ofwel gekoppeld aan het bestand). De datum geeft dan ook een versie aan. (2:64, 5:55, 11:40, 16:58)

Soms wordt de naamconventie opgelegd door de gebruikte software die altijd bestanden in een bepaald format aanmaakt (11:45)

Naamconventie groep

Naamconventie groep, md nmc groep			
Als naamconventies per groep worden gebruikt zijn er twee vormen	Genoemd in	#gen	#int
Een verplichte structuur waarin iedereen zijn datafiles met een vaste naamgevingsconventie moet zetten	2:40, 8:5, 8:31, 15:38, 1:48	5	4
Een conventie die door één onderzoeker wordt gehanteerd waarna anderen zijn/haar voorbeeld volgen, waarbij de methode van de onderzoeker als sjabloon wordt gebruikt	2:25, 12:4	2	2
Voorbeelden van quotes: 1:48: Interviewer: ……Geïnterviewde: I guess try to but the problem is I don't really do a lot of the research myself so it's best as I can try to get the students to do it. Naming conventions I definitely have. So I actually have a little naming file, naming conventions file that I share for every students they have project names and dates for naming but for documents I do have version control. …. 12:4: Geïnterviewde: And what's nice is that you can enter some meta data there. And what we tend to do it's become a kind of within the research group. We tend to automatically put a fairly descriptive file name on it. It's the kind of thing that you think one person starts doing and it's kind of spread. So automatic so you can set up a kind of template for the file name and then you just change your put in the individual. I can show it to you if you want to later. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 53, Naamconventie groep

Op basis van TABEL 52 en TABEL 53 lijkt er bij de groep geïnterviewden meer voorkeur te zijn voor een persoonlijke invulling dan voor een groepsinvulling. Naamconventies op groepsniveau lijken nauwelijks te zijn ingeburgerd. Ze zijn niet verplicht volgens het data management plan en onderzoekers volgen bij voorkeur hun eigen conventies (2:46, 6:48, 9:13).

11.3.6. Metadata inhoud

Uit het literatuuronderzoek volgde voor de inhoud van metadata: Bij voorkeur zouden Metadata tenminste moeten bestaan uit een dataset identifier, titel, beschrijving, de maker, contact (persoon/instituut), publicatiedatum, versie, identifiers van de makers en de licentie. Daar voegen de interviews weinig aan toe. Wel heeft men ideeën over de locatie van de metadata (TABEL 54).

Locaties van metadata, md inhoud			
Locaties van metadata	Genoemd in	#gen	#int
De publicatie zelf bevat de metadata door een verband te leggen tussen variabelen in de publicatie en de dataset	5:19, 5:20, 5:22, 5:29, 9:19	5	2
Metadata in losse documenten zoals Excel sheets	3:27, 3:28	2	1
Metadata in lablogboeken	3:28, 3:29, 3:30, 3:47	4	1
Voorbeelden van quotes: 9:19: Interviewer: Waar moet ik dan aan denken. De metadata? Wat beschrijven jullie in de metadata? Geïnterviewde: Nou ja wat in de data staat is meestal vrij kort, vaak 1 zin. Vaak een verwijzing naar het artikel: Die parameter van de vergelijking staat in de eerste kolom, een andere kolom is die parameter uit die vergelijking. Interviewer: En dan is je publicatie in feite je metadata. Geïnterviewde: In feite wel ja, Want daar staat het dan uitgelegd. 5:20: Geïnterviewde: Ik probeer altijd zoveel mogelijk de labels of de beschrijving van de labels te matchen met wat er uiteindelijk in de publicatie staat en ook dus met wat er in de vragenlijst staat zodat iedereen het ook terug kan vinden. Dat is misschien niet altijd het geval maar als je eenmaal gaat publiceren en je gaat die data delen. Dan zul je toch, dan zul je dat duidelijk moeten maken aan mensen wat bij wat hoort welke vragen uit je enquête hoort en bij welke variabele in je SPSS bestand. En hoe heb je dat dan uiteindelijk gebruikt voor je analyses in je papers. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 54, Locaties van metadata

Belangrijk om op te merken is het feit dat slechts drie geïnterviewden het over de locatie van metadata hebben. Er valt weinig te zeggen over het algemene gebruik van deze uitkomsten.

11.3.7. Enkele opmerkingen bij de uitkomsten uit de interviews aangaande metadata

T.a.v. veiligheid: soms is de data onleesbaar zonder de bijbehorende metadata. Dat zou als een vorm van beveiliging ingezet kunnen worden. Geïnterviewde 3 slaat de metadata apart op bij een extra beveiligde dienst (3:28, 3:29, 3:30, 3:47). Dit wordt slechts in één interview genoemd.

Standaardisatie kan voorkomen door een standaard te gebruiken als Dublin Core, of door het gebruik van voorbeelden uit de afdeling of sectie. Metadata kunnen ook gestandaardiseerd aangemaakt worden, zodat ze automatisch tijdens het verkrijgen van de onderzoeksdata worden geschreven (2:26, 3:38, 7:44, 8:3, 8:16, 12:4, 12:5, 12:6). In dat geval verzorgt de machine (bijvoorbeeld het meetsysteem) een standaard. Geïnterviewde 2 geeft aan dat dit een voorkeurswerkwijze zou moeten zijn, mits de implementatie eenvoudig is (2:22, 2:26). Dit wordt in vijf interviews besproken en lijkt daardoor wat sterker te zijn.

11.3.8. Code co-occurrence binnen de metadata categorie

De onderlinge verwantschappen binnen de metadata categorie staan in (TABEL 55).

Verwantschappen binnen de MD categrorie			
	Code 1	Code 2	Coëfficiënt
1	md definitie	md ja	0.40, 7
2	md fdst persoonlijk	md nmc persoonlijk	0.29, 6
3	md fdst groep	md nmc groep	0.24, 4
4	md ja	md vorm	0.23, 6

Uitleg:

1. Geïnterviewden die praten over hoe zij metadata in hun werk definiëren, geven hiermee vaak ook aan dat ze het überhaupt gebruiken. Deze verwantschap lijkt wel sterk te zijn. Mensen die zeggen metadata te gebruiken, hebben waarschijnlijk een beeld bij wat het is.
2. Naamconventies en folderstructuren worden vaak in één adem genoemd. Als de één persoonlijk is, dan is de ander dat meestal ook.
3. Naamconventies en folderstructuren worden vaak in één adem genoemd. Als de één een groepsrichtlijn volgt, dan doet de ander dat meestal ook.
4. Geïnterviewden die praten over in welke vorm ze metadata gebruiken, geven hiermee vaak (impliciet) ook aan dat ze het überhaupt gebruiken.

De sterkte van een verwantschap (co-occurrence) tussen twee codes wordt uitgelegd in Bijlage Uitleg Co-Occurrence/verwantschap.

Tabel 55, Verwantschappen binnen de MD categorie

De tabel geeft geen aanleiding voor de aanpassing van de eerdere conclusies rondom metadata. Het laat vooral de verwantschap van de verschillende onderdelen zien.

11.3.9. Metadata conclusie

De interviewers bevestigen de uitkomsten uit de literatuur, maar wel met enige kanttekeningen. De definitie wordt enigszins uitgebreid. Het gebruik van standaarden en domein specifieke standaarden blijkt bij deze geïnterviewden maar matig voor te komen. Het gebruik van naamconventies en folderstructuren wordt veel breder gedragen bij de geïnterviewde populatie. Hier blijkt dat er een verschil is tussen opgelegde, geadviseerde en vrije naamconventies en folderstructuren.

1. Definitie: Metadata zijn data die de onderzoeksdata leesbaar, interpreteerbaar en bruikbaar te maken voor anderen. Bovendien kunnen metadata een indicatie geven van de waarde van de verzamelde gegevens en daarmee mogelijk een juiste herhaling van het experiment bevorderen (validiteit).
2. Vorm: Metadata kunnen in lagen geformuleerd worden. De eerste laag is een breed toepasbare standaard (zoals de Dublin Core standaard voor metadata (DCMI 2018)). Deze wordt gevolgd door een domein/discipline specifieke laag en vervolledigd met een laag die de data op folder-, file- en naamniveau beschrijft. Het beschrijven van de data op folder- en naamniveau kan een opgelegde structuur zijn, een geadviseerde of een volledig vrije.
3. Inhoud: Metadata zouden tenminste moeten bestaan uit een dataset identifier, titel, beschrijving, de maker, contact (persoon/instituut), publicatiedatum, versie, identifiers van de makers en de licentie.

Waar men gedurende het onderzoek metadata in bijhoudt is niet algemeen te zeggen. Drie geïnterviewden noemen hier lablogboeken, losse documenten en/of de publicatie zelf.

Machinaal geschreven metadata (bijvoorbeeld door de meetopstelling) is per definitie gestandaardiseerd voor het onderzoek waarin het wordt gebruikt. Vijf geïnterviewden gebruiken een dergelijke opzet en één van hen noemt het de voorkeur.

11.4. Datadocumentatie

Voor datadocumentatie is tijdens het literatuuronderzoek geen definitie opgesteld. In lijn met de structuur van de paragrafen over onderzoeksdata en metadata, wordt ook voor datadocumentatie van een definitie uitgegaan.

De case organisatie en in het bijzonder het 4TU datacenter, geeft richtlijnen voor datadocumentatie (Anon n.d.). Deze richtlijnen worden samengevat in de definitie die de WUR (één van de 4 TU's) op haar website geeft (Anon n.d.). Hieronder staat de definitie vertaald naar het Nederlands:

"Datadocumentatie tijdens onderzoek betekent het georganiseerd bijhouden van aantekeningen over hoe de data zijn verzameld, wat de resulterende databestanden zijn en hoe ze zijn verwerkt."

De WUR voegt er voorwaarden aan toe om te controleren of data documentatie voldoet:

"Uw documentatie moet de volgende vragen beantwoorden:

1. Wat bevat mijn dataset? Welke afkortingen zijn gebruikt en wat betekenen ze? Hoeveel data zijn verzameld? Welke software is nodig om ze te lezen?
2. Hoe is mijn dataset verzameld? Wie heeft de gegevens verzameld? Op welke data zijn ze verzameld?
3. Hoe zijn mijn gegevens verwerkt?"

Verder beschrijft men hoe de datadocumentatie kan worden bijgehouden.

Met andere woorden, voor het gebruik van datadocumentatie zijn 4 zaken van belang:

1. De data zelf en hoe ze kunnen worden gelezen;
2. Hoe de data zijn verzameld;
3. Hoe de data zijn bewerkt;
4. Hoe en waar(in) de datadocumentatie wordt bijgehouden.
5. In de interviews wordt ook gesproken over data linken, dit is het verwijzen naar een dataset in bijvoorbeeld een ELN, waarbij je door het volgen van de verwijzing de dataset kunt vinden (de verwijzing kan bijvoorbeeld een hyperlink zijn).

De subcodes staan opgesomd in onderstaande (TABEL 56).

Subcodes in de codecategorie datadocumentatie	
Subcode omschrijving	Subcode in Atlas
De data zelf en hoe ze kunnen worden gelezen	dd inhoud
Hoe de data zijn verzameld	dd data verzamelen
Hoe de data zijn bewerkt	dd data bewerken
Hoe en waar(in) de datadocumentatie wordt bijgehouden	dd papier, dd readme, dd eln
Data linken	dd link

Tabel 56, Subcodes in de codecategorie datadocumentatie

11.4.1. Datadocumentatie inhoud

Wat er in datadocumentatie kan staan, wordt beschreven in TABEL 57.

Datadocumentatie inhoud, dd inhoud			
Datadocumentatie inhoud	Genoemd in	#gen	#int
Datadocumentatie is vaak een kort stukje geschreven tekst, waarin de data worden beschreven	2:57, 5:21, 5:28, 5:63, 8:18, 11:26, 11:47, 16:31	8	5
Soms is men specifieker door:			
Het experiment (en/of de omstandigheden er omheen) te beschrijven	2:49, 11:47, 12:2	3	3
De bewerkingsstappen (of het protocol) te beschrijven	1:55, 2:49, 2:51, 10:16, 11:47	5	4
De uitkomst te beschrijven	2:49, 2:51, 11:47, 12:41	4	3
Stap voor stap te beschrijven wat er met de data gebeurt (de transformatie van ruwe data naar de figuren in de publicatie)	3:37, 3:39, 5:30, 10:16, 15:42	5	4
Voorbeelden van quotes: 2:49: Geïnterviewde: Then I print that photo and I glue it to my lab journal and then I write and I explain. OK. This is experiment. These are my samples. This is a protocol I've done and this is my outcome. 10:16: Geïnterviewde: Ja daar gaat die provenance echt een rol spelen. Interviewer: En hoe pak je dat dan aan? Hoe zorg je dat het volgbaar blijft? Interviewer: Wat ik dus probeer te doen is gewoon Notebooks, dus met Jupyter notebooks werk ik veel. Daar probeer ik dan heel netjes die provenance in vast te leggen. Ik pak mijn dataset met de experimenten, ik doe deze, deze en deze processing steps en dan schrijf ik het weg naar een nieuwe date file met een andere naamgeving. En dan kan iedereen dus zien, ok, Dit zijn stappen die uitgevoerd zijn. Dat is hoe ik het probeer te doen. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 57, datadocumentatie inhoud

Twee specifieke voorbeelden om de data leesbaar te maken voor anderen, worden 11:5 en 11:26 gegeven. Men levert het script mee waarmee de bewerking verricht is. Door het script te lezen, kun je de data begrijpen (11:5). Men levert een programmaatje mee waarmee de data wordt uitgelegd (11:26). Dit lijken wel uitzonderingen te zijn.

In 12:7 wordt aangegeven dat er geen aparte aandacht wordt besteed aan datadocumentatie, omdat alles in het ELN staat.

11.4.2. Datadocumentatie over data verzamelen

Er wordt in de interviews niet ingegaan op het daadwerkelijk beschrijven van de verzameling van data in datadocumentatie. Maar door het experiment te beschrijven en stap voor stap te beschrijven wat er met de data gebeurt (zie TABEL 57), lijkt het voor de hand te liggen dat men ook de dataverzameling i.i.g. ten dele beschrijft.

11.4.3. Datadocumentatie over data bewerken

Uit de eerste paragraaf (TABEL 57) kan worden overgenomen:

- De bewerkingsstappen (of het protocol) te beschrijven 1:55, 2:49, 2:51, 10:16, 11:47;
- Stap voor stap te beschrijven wat er met de data gebeurt (de transformatie van ruwe data naar de figuren in de publicatie) 3:37, 3:39, 5:30, 10:16, 15:42.

11.4.4. Datadocumentatie over medium waarin/waarop het wordt bijgehouden

Er zijn verschillende plekken waar datadocumentatie kan worden bijgehouden (TABEL 58).

Datadocumentatie wordt op diverse manieren bijgehouden, dd papier, dd readme, dd eln			
Manieren om datadocumentatie bij te houden	Genoemd in	#gen	#int
In een readme file (soms met een andere naam). Dit kan een plat tekst document zijn, of een bewerkt tekstdocument in bijvoorbeeld Word, PowerPoint of een Excelsheet	2:16, 2:21, 5:62, 6:11, 6:53, 7:19, 10:13, 11:47, 15:42	9	7
In een ELN	1:5, 1:56, 2:21, 2:57, 3:37, 3:39, 6:34, 10:16, 10:19, 12:1, 12:10, 12:11, 12:36, 12:41, 15:42	15	7
Op papier	2:16, 2:49, 2:51, 2:57, 6:2, 6:11, 6:12, 6:34, 15:42	9	3
Geautomatiseerd vanuit de software van het systeem waarmee wordt gemeten of gerekend	2:6, 3:39, 7;60	3	3
Voorbeelden van quotes: 2:16: Geïnterviewde: So there are people who have handwritten notes or lab journals and there are people who are typing in Word or and there are also some people using OneNote and they seem to be happy with it 15:42: Interviewer: En datgene wat met datasets wordt gedaan, waar wordt dat bijgehouden? Houden mensen lab notebooks bij? Geïnterviewde: Ja, soms digitaal soms op schrift en soms niet. Dus alle drie. Digitaal, daar adviseren wij niet in. We hebben geen vaste manier waarop dat gedaan wordt dus iedereen doet dat op zijn eigen manier. Ik heb gezien dat mensen daarvoor notitieprogramma's gebruiken of inderdaad spreadsheets voor gebruiken. Sommige mensen een readmefile die bij de data staat waar dan beschreven staat wat er gedaan is met die data. En wat ik al zei: op schrift er worden logboeken bijgehouden. Wordt veel gedaan, trouwens. Veel logboeken, analoge logboeken. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS)			

Tabel 58, Datadocumentatie wordt op diverse manieren bijgehouden

Een reden om een papieren lablogboek te gebruiken kan zijn dat in sommige laboratoria geen elektronica is toegestaan (6:34).

11.4.5. Datadocumentatie, het linken naar datasets

Soms hebben mensen maatregelen genomen (of is het een wens) om in hun datadocumentatie te kunnen verwijzen naar de dataset waar over geschreven wordt. Een dergelijke link kan een klikbare link zijn (zoals een hyperlink) of een geschreven verwijzing. 1:55, 1:56, 1:57, 3:40, 7:58, 10:19, 12:36, 14:33 (acht quotes in zes interviews).

11.4.6. Code co-occurrence binnen de datadocumentatie categorie

Hoe de verschillende codes van datadocumentatie samenhangen staat beschreven in (TABEL 59).

Verwantschappen binnen de atadocumentatie categorie			
	Code 1	Code 2	Coëfficiënt
1	dd data bewerken	dd inhoud	0.29, 4
2	dd data bewerken	dd eln	0.26, 4
3	dd eln	dd inhoud	0.25, 5
4	dd eln	dd link	0.21, 3
Uitleg: De sterkste verwantschap bestaat tussen de drie codes datadocumentatie ELN, datadocumentatie Inhoud en datadocumentatie data bewerken. Een mogelijke reden kan zijn dat een lablogboek in de regel gedurende het hele onderzoek wordt bijgehouden. De kans dat alle data gerelateerde activiteiten er in staan is daarmee wellicht wat groter dan bij losse bestanden zoals een readme file. De inhoud overlapt in de beschrijving sowieso behoorlijk met bewerken. Het verband tussen dd eln en dd link wordt maar in drie interviews benoemd, omdat het ook weinig gangbaar is (de combinatie ELN gebruiken en naar data linken). De sterkte van een verwantschap (co-occurrence) tussen twee codes wordt uitgelegd in BIJLAGE UITLEG CO-OCCURRENCE/VERWANTSCHAP.			

Tabel 59, Verwantschappen binnen de datadocumentatie categorie

Bovenstaande suggereert dat gebruik maken van een elektronische lablogboek, de meest volledige beschrijving van de data activiteiten in de hand werkt.

11.4.7. Datadocumentatie conclusie

De definitie van datadocumentatie is:
Datadocumentatie tijdens onderzoek betekent het georganiseerd bijhouden van aantekeningen over hoe de [onderzoeks]data zijn verzameld, wat de resulterende databestanden zijn en hoe ze zijn verwerkt.

Datadocumentatie wordt op diverse manieren bijgehouden:

- In een readme file (soms met een andere naam). Dit kan een plat tekst document zijn, of een bewerkt tekstdocument in bijvoorbeeld Word, PowerPoint of een Excelsheet;
- In een ELN;
- Op papier;
- Geautomatiseerd vanuit de software van het systeem waarmee wordt gemeten of gerekend.

Het komt voor dat men in de datadocumentatie een link (verwijzing) opneemt naar de beschreven data (dat kan een klikbare link zijn of een geschreven verwijzing).

Mensen die en lablogboek gebruiken lijken vaker inhoudelijke beschrijvingen te maken van de onderzoeksstappen, vaker aandacht te besteden aan het bewerken van de data in de onderzoeksstappen en vaker naar datasets te linken.

11.5. Data opslag

Het uitgangspunt van dit onderzoek is dat onderzoeksdata op veel plekken worden opgeslagen en dat het voor de universiteit of voor een onderzoeker lastig is om alle data die bij een onderzoek horen te vinden of alle data van één onderzoeker. Dat geldt dan wel voor de tijd dat het onderzoek loopt. Na het onderzoek wordt er gearchiveerd en is het vinden van de data een stuk minder lastig. Hieronder wordt uitgewerkt op wat voor verschillende media de data staan, waar die media aan moeten voldoen als het gaat om capaciteit en betrouwbaarheid en wordt de bevestiging gezocht voor het feit dat data verspreid over verschillende media opgeslagen staan. In de interviews komt dat tot uitdrukking in de codes die staan benoemd in (TABEL 60).

Subcodes in de codecategorie data opslag	
Subcode omschrijving	Subcode in Atlas
On premise gecentraliseerde opslag	opsl centraal
Opslag bij share and sync oplossingen	opsl share and sync
Opslag externe datahouder	opsl ext datahouder
Opslag voor code	opsl code
Opslag overal	opsl overal
Opslag capaciteit	opsl capaciteit
Opslag betrouwbaarheid	opsl betrouwbaarheid

Tabel 60, Subcodes in de codecategorie data opslag

11.5.1. Data opslag centraal

De case organisatie biedt verschillende mogelijkheden om data op centrale locaties (netwerkshares op on premise opslagsystemen) op te slaan. Bij on premise gecentraliseerde opslag worden met name de verschillende soorten netwerkshares genoemd (TABEL 61).

Data opslag centraal			
On premise netwerkshares	Genoemd in	#gen	#int
De projectshare, een gedeelde netwerkopslagplaats die voor een specifiek onderzoeksproject wordt aangemaakt	1:26, 2:35, 3:8, 3:43, 3:57, 6:41, 14:11, 14,15, 14:16, 15:22, 15:28	11	6
Bulk, een gedeelde netwerkopslagplaats die gebruikt kan worden voor grote hoeveelheden data (wordt vervangen door de projectshare)	10:22, 15:21	2	2
Groupshares (niet specifiek voor onderzoeksdata bedoeld, maar meer voor het delen van data in een groep om andere redenen)	15:21, 16:26	2	2
Voorbeelden van quotes: 3:8: Geïnterviewde: En dan bedoel [je] shares: de bulk, de project store? Geïnterviewde: Vooral de project store 6:41: Geïnterviewde: Het is heel erg persoonlijk, op persoonlijke computers, persoonlijke laptop heel erg Dropbox. Af en toe Googledrive kom ik tegen af en toe Surfdrive. Maar ook projectdrive. Het gaat wel steeds meer die kant uit dat mensen daar wat meer bewust van zijn. Als je daar thuis toegang toe wil hebben, is het gewoon een stuk ingewikkelder dan Dropbox en Google en al die zaken. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 61, Data opslag centraal

De aantallen suggereren dat bij het gebruik van centraal aangeboden netwerkshares, met name de projectshare wordt gebruikt. Geïnterviewde 16 geeft aan dat er ook gebruik wordt gemaakt van een wiki en een repository die bedoeld is om afgeronde onderzoeken op te slaan (16:19, 16:20 en 16:64). Die laatste valt buiten de scope van het onderzoek en de Wiki is dermate specifiek en maar één keer benoemd, dat hij niet wordt meegenomen naar de conclusies.

Eén geïnterviewde geeft aan de data in de mail op te slaan. Op die manier is het doorzoekbaar en is het eenvoudig te gebruiken als takenlijst (de mails worden dan als taken beschouwd). Het gaat hier om kleine hoeveelheden data (5:41, 5:42, 5:43). Ook dit is te specifiek om mee te nemen naar de conclusies.

11.5.2. Data opslag bij share and sync oplossingen

Bij een share and sync oplossing wordt de data op een lokaal systeem opgeslagen en vervolgens automatisch naar een centraal (cloud) systeem gesynchroniseerd (men kan per folder instellen of deze wel of niet gesynchroniseerd moet worden). Data kan gedeeld worden met mensen binnen en buiten de case organisatie. Bekende commerciële voorbeelden zijn Dropbox en Google Drive. Binnen de Nederlandse academische gemeenschap wordt ook gebruik gemaakt van SurfDrive (een niet commerciële oplossing aangeboden door Surf). Genoemde voorbeelden staan samengevat in TABEL 62.

Data opslag Share and Sync			
Genoemde vormen	Genoemd in	#gen	#int
Dropbox	1:25, 1:28, 1:35, 1:38, 1:40, 2:10, 2:11, 2:12, 2:35, 3:10, 6:41, 10:1, 10:8, 11:21, 14:11, 14:15, 14:41, 15:30, 16:65	19	9
GoogleDrive	1:28, 2:11, 6:41	3	3
SurfDrive	2:32, 3:52, 6:17, 6:41, 7:69, 16:65	6	5
Voorbeelden van quotes: 1:25: Geïnterviewde: And for documents and data, basically we use Dropbox. 14:41: Geïnterviewde: Maar mensen gebruiken allemaal hun eigen Dropbox. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 62, Data opslag Share and Sync

Share and Sync oplossingen worden door tien verschillende geïnterviewden gebruikt. Dropbox lijkt hierbij vaker te worden gebruikt dan GoogleDrive en SurfDrive.

11.5.3. Data opslag externe datahouder

Externe datahouders: Alle decentrale devices en of systemen (en mogelijk hun backups naar lokale media) waar onderzoeksdata op verzameld worden (ELN's, HPC omgevingen, meetinstrumenten, laptops, desktops, flashdrives enzovoort). Voorbeelden die in de interviews genoemd worden staan in TABEL 63.

Data opslag Onderzoeksdatahouder			
Genoemde vormen	Genoemd in	#gen	#int
HPC	1:38, 1:40, 8:6, 9:30, 11:16, 14:10, 14:29, 15:28	8	6
Meetsystemen	2:47, 2:53, 3:21, 7:63, 12:14	5	4
Eigen systemen in de afdeling, veelal virtuele systemen, die in de afdeling worden gebruikt	7:42, 8:31, 8:34, 16:16, 16:17, 16:20, 16:22, 16:23, 16:25, 16:26, 16:36	11	3
Laptops/desktops	2:1, 2:9, 2:55, 3:9, 3:18, 3:21, 6:17, 8:6, 9:24, 10:8, 12:14, 12:19, 16:30	13	8
USB sticks	2:53, 15:5, 15:30	3	2
Externe harde schijven	2:1, 2:55, 3:50, 6:57, 11:13, 12:22, 15:5	7	6
Voorbeelden van quotes: 12:14: Geïnterviewde: So where the data is all on the computer that runs experiments so all the kind of raw data. Then everything else is on my computer. 14:10: Interviewer: En moet je ook in staat zijn om back ups en restores te maken? Of is het zo van je maakt voldoende stappen in zo'n onderzoek dat als je een keer een stap verliest dat dat niet erg is? Geïnterviewde: Nou kijk, het gebeurt natuurlijk niet vaak. Dan zouden we wel weer kunnen hergenereren. Deze spullen staan allemaal op de hpc's daar wordt volgens mij een back up van gedraaid want de meeste stukken tenminste op sommige stukken. Geïnterviewde: Maar ik hamer er bij iedereen op dat ze hun eigen spullen netjes weer ergens moeten zetten want.....eh ja. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 63, Data opslag Onderzoeksdatahouder

Voor vrijwel al deze datahouders geldt dat ze eigendom zijn van de onderzoeker of de onderzoeksafdeling.

11.5.4. Data opslag voor code

De scripts spelen vaak en belangrijke rol in het onderzoek, zoals bij de beschrijving van onderzoeksdata al is aangegeven. Er worden ook specifieke gespecialiseerde opslagdiensten voor gebruikt (TABEL 64).

Data opslag voor code			
Genoemde vormen	Genoemd in	#gen	#int
Git Hub	9:10, 10:9, 10:10, 15:32, 15:36, 16:34, 16:35	7	4
Git Lab	3:11, 15:36	2	2
Bit Bucket	1:24, 14:32, 16:34, 16:35	4	3
Geen duidelijke benoeming, maar wel een specifieke code opslag (Git genoemd i.p.v. Git Hub of Git Lab, ander soort version control)	1:24, 2:18, 3:10, 10:7, 10:8, 10:20, 10:21, 11:10, 11:12, 11:13, 11:18	11	5
Voorbeelden van quotes: 2:18: Geïnterviewde: One thing I've found out is that with programming many researchers like to use Git. 15:32: Geïnterviewde: Voor de nieuwere versies daarvan gaan we gebruikmaken van GitHub. Daar is echt een complete rewrite van de software gemaakt, van de library. Die is van C overgezet naar C++ en die zal niet meer bij de case organisatie neergezet worden om op te halen die wordt via GitHub gedistribueerd. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 64, Data opslag voor code

11.5.5. Data opslag overal

De voorgaande paragrafen geven goed weer data onderzoeksdata overal staan opgeslagen. Dat dit problemen geeft in de vindbaarheid van de data wordt nog eens expliciet aangegeven door enkele geïnterviewden (1:34, 1:35, 7:1, 7:58, 8:30, 10:1, 15:23, 16:1; acht keer genoemd door zes verschillende geïnterviewden).

11.5.6. Data opslag capaciteit

De hoeveelheid opslag die men nodig heeft, wisselt sterk per onderzoeker. In de interviews worden getallen genoemd van enkele Mb's (5:4) tot snapshots van 40Tb (11:7). In dat laatste geval, zou je zelfs enkele daarvan per minuut kunnen maken, maar je zou niet de capaciteit hebben om het op te slaan (11:8). Op dat moment gaat men pragmatisch kijken naar het niveau van de data, wat is precies van belang om op te slaan, zodat anderen het onderzoek kunnen verifiëren, repliceren of de data voor verder onderzoek kunnen gebruiken (7:29, 7:39, 9:32, 11:8) (zie ook Archiveerbaarheid).

11.5.7. Data opslag betrouwbaarheid

Er wordt weinig gesproken over betrouwbaarheid. Men vindt opslag betrouwbaar als:

- Er backups worden gemaakt (5:47, 6:19, 8:37; drie keer genoemd in drie interviews);
- Als alleen mensen er bij kunnen die zijn geautoriseerd (3:43, 6:19; twee keer genoemd in twee interviews).

Deze beide aspecten komen bij beveiliging uitgebreider terug. Hoewel geen van de geïnterviewden de link legt, lijkt het aannemelijk dat zij vinden dat betrouwbaarheid afhangt van een goede beveiliging.

11.5.8. Code Co-occurrence binnen de data opslag categorie

Er bestaat geen grote verwantschap tussen de verschillende opslaglocaties. Dit zou wellicht suggereren dat men zich bij voorkeur tot één type opslag beperkt of dat onderzoekers ieder unieke

combinaties van verschillende opslagvormen gebruiken. Dat zou verder onderzocht moeten worden en kan niet uit de beschikbare data geconcludeerd worden.
Er is ook geen sterke verwantschap tussen een bepaald type opslag en de betrouwbaarheid of capaciteit ervan.

11.5.9. Data opslag conclusie

Relevante data voor het lopende onderzoek staan op diverse plekken opgeslagen. Dat kan problemen geven als het gaat om de vindbaarheid van die data. Locaties waar data staan opgeslagen zijn:

- Centraal op TU systemen bedoeld voor de gehele campus, de zogenaamde netwerkshares. Hierbij bestaat een grote voorkeur voor de projectshare. Er worden ook afwijkende locaties genoemd, maar die lijken dan heel specifiek voor één onderzoeker te zijn.
- Share and Sync oplossingen worden door de meeste geïnterviewden gebruikt. Dropbox lijkt hierbij vaker te worden gebruikt dan GoogleDrive en SurfDrive.
- Externe onderzoeksdatahouders vrijwel altijd eigendom van de onderzoeker of onderzoeksafdeling:
    - HPC;
    - Meetsysteem;
    - Eigen systemen in de afdeling, veelal virtuele systemen die in de afdeling worden gebruikt;
    - Laptops/desktops;
    - USB stick;
    - Externe harddisks.
- Voor code worden veelal specifieke opslagsystemen gebruikt. Genoemd werden:
    - Git Hub;
    - Git Lab;
    - Bit Bucket;
    - Geen duidelijke benoeming, maar wel een specifieke code opslag (Git genoemd i.p.v. Git Hub of Git Lab, ander soort version control).

Opslag wordt als betrouwbaar beschouwd als deze goed beveiligd is (DATA BEVEILIGING).
De behoefte aan capaciteit wisselt sterk per onderzoek.

11.6. Data delen

Op basis van de conclusies uit de interviews voor data, metadata en datadocumentatie, kan data delen in deze context als volgt gedefinieerd worden: Data delen is het beschikbaar stellen van research data, met de bijbehorende metadata en/of datadocumentatie aan anderen.

Uit de interviews komt daar het volgende uit naar voren (TABEL 65):

Subcodes in de codecategorie data delen	
Subcode omschrijving	Subcode in Atlas
Hoe men data deelt, vanaf een type systeem/opslagmedium	del centraal, del datahouders, del mail, del sync and share
Met wie men data deelt, data delen binnen de TU of erbuiten	del intern, del extern
Waarom men al dan niet data wil delen	del motivatie

Tabel 65, Subcodes in de codecategorie data delen

11.6.1. Data delen hoe

Elf van de veertien geïnterviewden geven verschillende opties om data met anderen te kunnen delen (TABEL 66).

Data delen hoe: del centraal, del datahouders, del mail, del sync and share			
Genoemde vormen	Genoemd in	#gen	#int
Vanaf een centraal systeem zie ook Data opslag centraal)	1:26, 2:33, 2:34, 2:35, 2:36, 5:15, 5:16, 5:51, 5:52, 6:15, 6:37, 6:38, 6:39, 9:26, 11:28, 11:30, 14:15, 14:16, 15:25, 15:31, 16:8, 16:38, 16:40, 16:41, 16:42, 16:64	26	9
Vanaf externe datahouders (zie ook Data opslag externe datahouder			
HPC	11:29, 14:11	1	1
Externe harde schijven en usb sticks	1:24, 2:18, 3:10, 10:7, 10:8, 10:20, 10:21, 11:10, 11:12, 11:13, 11:18	11	5
Via verschillende Share and Sync oplossingen zoals benoemd in Data opslag bij share and sync oplossingen, maar ook WeTransfer en SurfFilesender	1:27, 2:33, 2:35, 2:36, 5:15, 5:16, 6:38, 6:39, 8:29, 9:27, 14:12, 14:15, 15:25	13	8
Vanaf systemen die helpen met code beheer (zie ook Data opslag voor code)	1:24, 1:27, 8:27, 10:9, 10:10, 11:10, 15:32, 15:36, 16:34, 16:35	10	6
Voorbeelden van quotes: 5:16: Interviewer: En met wie moet je delen? Geïnterviewde: SharePoint is binnen de case organisatie, maar Dropbox en Google documenten is eigenlijk altijd binnen Europese projecten, dus partners binnen Europa. Interviewer: En ook binnen de EU? Geïnterviewde: Ja altijd binnen de EU. 11:10 Geïnterviewde: Yeah I mean...The software is managed in a Git Repository so that we can go back and forth and have everything documented. So I mean it's also the sharing part. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 66, hoe data gedeeld worden

Op basis van de genoemde systemen, kan geconcludeerd worden dat data delen op twee manieren plaatsvindt:

- Door het op de eigen systemen te laten staan en derden toegang te geven (de persoon met wie gedeeld wordt, kan natuurlijk wel zelf een kopie maken);
- Door de data daadwerkelijk te kopiëren naar een systeem van de degene met wie gedeeld wordt.

11.6.2. Data delen met wie

Data worden gedeeld met mensen binnen en buiten de case organisatie. Binnen betekent binnen de groep tot binnen de universiteit (over faculteiten heen) en alles er tussen. Buiten de universiteit betekent (geografisch) met de US, andere EU staten en/of wereldwijd (niet specifiek benoemd) (TABEL 67).

Data delen met wie, del intern, del extern			
Waar bevinden de personen met wie gedeeld wordt zich	Genoemd in	#gen	#int
Intern	3:7, 5:16, 6:14, 8:29, 11:28, 11:29, 14:11, 15:29	8	7
Extern	1:27, 3:6, 3:21, 5:16, 5:17, 8:29, 11:30, 14:11, 14:14, 15:29, 15:31	11	7
Voorbeelden van quotes: 11:29: Interviewer: Then you're shared with the people in this building? Geïnterviewde: Depends. I mean generally everyone with access to HPC12 could get access. Interviewer: Is that only people from TU or is that people from outside as well? Geïnterviewde: That should be almost exclusively people from within. 15:31: Interviewer: Maar ook met partijen buiten de EU? Geïnterviewde: Ook wel. Ik probeer even te verzinnen. Er is één heel bekend voorbeeld wat ik misschien al eerder genoemd heb. De groep computational Imaging heeft al een jaar of dertig een library ontwikkeld en die wordt wereldwijd beschikbaar gesteld en die wordt nog steeds opgehaald en gebruikt door alle landen die er zijn. Interviewer: En dat is een website? Geïnterviewde: We hebben een website voor en we gebruiken onze eigen FTP server die we binnen de case organisatie hebben neergezet. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 67, Data delen met wie

11.6.3. Data delen waarom

Er zijn verschillende redenen om data te delen (TABEL 68).

Redenen om data te delen, del motivatie			
De belangrijkste redenen om data te delen zijn omdat:	Genoemd in	#gen	#int
Men er zelf wat aan heeft (de verwachting ook zelf data te krijgen, erkenning door peers en/of werkgever);	5:36, 5:37, 5:38, 5:66, 11:37	5	2
Men met anderen samenwerkt	6:13, 6:16, 6:64, 6:65, 12:13, 16:31	6	3
Het eenvoudig is om te doen	1:30, 3:53	2	2
Voorbeelden van quotes: 5:37: Geïnterviewde: dus institutioneel gezien zou je onderzoekers beter willen belonen, bijvoorbeeld tijdens hun jaarlijkse evaluatie. Wil je ervoor zorgen dat zij ook worden geëvalueerd op wat heb je bijgedragen aan delen met jouw community? Wat heb je bijgedragen aan open science? Dat dat één van de onderwerpen is, die standaard wordt besproken. Daar word je ook op geëvalueerd. 11:37: Interviewer: It's not something that's important for you? If it would not be there.... Geïnterviewde: It's not important, sometimes it's interesting to see that one paper is downloaded a lot. But that's not so important. What's more important is when they cite it, or use it, what they say about it. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 68, Redenen om data te delen

11.6.4. Redenen om niet te delen

Er zijn verschillende redenen waarom men data niet zou willen delen (TABEL 69).

Redenen om data NIET te delen, del motivatie			
De belangrijkste redenen om data niet te delen zijn:	Genoemd in	#gen	#int
Competitive advantage behouden	5:65, 5:67, 9:9, 14:21	4	3
Fysiek moeilijk (bijvoorbeeld hele grote dataset)	10:10, 16:43, 16:44	3	2
Moet veel ondersteuning bieden na delen	9:9, 16:31, 16:44	3	2
Het is geen eigen data (er staat bijvoorbeeld een licentie op)	10:10, 12:8, 12:9, 12:24, 12:28	5	2
Voorbeelden van quotes: 10:10: Geïnterviewde: Dat het kan zijn dat die data het resultaat is van een aantal processing steps op een heel grote dataset, kan betekenen dat we die grote dataset niet plaatsen op GitHub. En dat heeft dan weer te maken met die big store. Het andere probleem heeft er vaak mee te maken, dat ik die data gegenereerd heb met simulatiemodellen waarvan ik de onderliggende data ook niet altijd kan delen. Dus dan is het vaak dat je zegt: Ok, Ik gebruik dit simulatie model zie dit paper en daar moet ik stoppen. Het is niet aan mij om dit model te delen met anderen. 14:21: Geïnterviewde: Omdat ik niet wilde dat....eh ik wilde geen open. Ik wilde die code niet open hebben. Dat is mijn competitive advantage. Dus ik wil dat niet open van mij. Interviewer: Wat bedoel je met competitive advantage? Geïnterviewde: Nou er zijn heel veel andere wetenschappers die vanalles doen en er zit heel veel tijd in die codes, dus die ga ik niet zomaar aan iedereen geven. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 69, Redenen om data niet te delen

De twee middelste redenen uit TABEL 69 benadrukken dat men niet wil delen omdat niet laagdrempelig is en er mogelijk veel extra werk van komt. Dat suggereert dat eenvoudig kunnen delen betekent: Een laagdrempelige methode om te delen zonder nawerk.

11.6.5. Data delen datatransfer

In sommige gevallen is de hoeveelheid data zo groot, dat kopiëren via internet (of het TU netwerk) dermate lang duurt dat men alternatieven zoekt. Dat betekent dan de data halen op enkele externe

hard disks. 3:5, 3:7, 3:18 (fiets op de campus), 3:19, 3:20, 3:21 (trein naar België), 3:50, 15:5 (hard disks op laten sturen) (acht quotes van twee geïnterviewden).

11.6.6. Code Co-Occurrence binnen de data delen categorie

De onderlinge verwantschappen voor data delen staan in TABEL 70.

Verwantschappen binnen de data delen categorie			
	Code 1	Code 2	Coëfficiënt
1	del intern	del extern	0.36, 4
2	del centraal	del sync and share	0.23, 7

Uitleg:

1. Data intern en extern delen hebben een sterke verwantschap. Dat komt vermoedelijk doordat de bereidheid om data te delen sterker is dan of dat intern en/of extern gedeeld wordt.
2. Bij sync and share wordt decentraal opgeslagen data gesynchroniseerd met een centrale opslagplaats. We gaat het hier om verschillende centrale opslagplaatsen. Die voor sync and share is in de cloud en die voor centraal is on premisse.

Tabel 70, Verwantschappen binnen de data delen categorie

Het lijkt er niet op dat de co-occurrence extra informatie oplevert t.o.v. het voorgaande.

11.6.7. Data delen conclusie

Het lijkt er op dat de meeste TU onderzoekers data delen.
Zij doen dat omdat:

- Men er zelf wat aan heeft (de verwachting ook zelf data te krijgen, erkenning door peers en/of werkgever);
- Men met andere samenwerkt;
- Het eenvoudig is om te doen.

Zij delen via:

- het op de eigen systemen te laten staan en derden toegang te geven (de persoon met wie gedeeld wordt, kan natuurlijk wel zelf een kopie maken);
- door de data daadwerkelijk te kopiëren naar een systeem van de degene met wie gedeeld wordt;
- Fysiek transport van de datahouder.

Data worden gedeeld met mensen binnen en buiten de case organisatie

- Intern, binnen de case organisatie (van binnen de groep tot binnen de universiteit):
- Extern, buiten de case organisatie met als belangrijke grenzen binnen en buiten de EU.

Als men er voor kiest om niet te delen is dat omdat:

- Men wil competitive advantage behouden;
- Het fysiek moeilijk is (bijvoorbeeld hele grote dataset, ondersteunde software);
- Er veel ondersteuning wordt gevraagd na het delen;
- Het geen eigen data is (er staat bijvoorbeeld een licentie op).

Het lijkt er op dat men eerder geneigd is om data te delen als de manier waarop laagdrempelig is en als het verder weinig nawerk oplevert.

11.7. Data beveiliging

Data beveiliging is een container begrip (eigenlijk een samenvoeging van twee containerbegrippen) en er zijn diverse definities voor te vinden. NIST definieert data beveiliging als (vertaald): “Bescherming van data tegen ongeautoriseerde (per ongeluk of opzettelijk) wijziging, vernietiging of openbaarmaking.” (Kissel, Locke, and Gallagher, p59, 2011). Verschillende aspecten die samen of alleen tot data beveiliging leiden, komen in de interviews naar voren. Deze aspecten staan in TABEL 71 samen met hun bijbehorende codes. Enigszins gerelateerd aan beveiliging is verder logging. Logging kan helpen in de beveiliging, maar kan ook breder ingezet worden.

Subcodes in de codecategorie data beveiliging	
Subcode omschrijving	Subcode in Atlas
Authenticatie	bvlg authenticatie
Autorisatie	bvlg autorisatie
Encryptie	bvlg encryptie
Backup (en restore)	bvlg backup
Contract (alleen toegang voor gecontracteerde derden)	bvlg contract
Logging	logging

Tabel 71, Subcodes in de codecategorie data beveiliging

Authenticatie, autorisatie en encryptie zijn handelingen op het opslagsysteem om de (toegang tot) de data te beveiligen (TABEL 72). Backup is bedoeld om op terug te kunnen vallen in geval van een calamiteit (m.b.v. een restore, TABEL 73) en contracten zijn redenen om een bepaalde mate van beveiliging toe te passen. De paragrafen hieronder worden analoog hieraan ingedeeld.

11.7.1. Data beveiliging authenticatie, autorisatie en encryptie

Data beveiligen met authenticatie, autorisatie en encryptie, bvlg authenticatie, bvlg autorisatie en bvlg encryptie			
De belangrijkste redenen om data niet te delen zijn:	Genoemd in	#gen	#int
Authenticatie, weten wie er toegang heeft	1:42, 3:17, 3:43, 3:44, 5:48, 5:52, 6:19, 6:23, 8:11, 8:12, 8:15, 8:28, 14:23, 14:26, 14:27, 14:34, 16:51	17	7
Autorisatie, de juiste toegang voor ene vertrouwde persoon (bijvoorbeeld het recht om te lezen)	1:42, 3:17, 3:53, 3:54, 5:48, 5:52, 6:19, 6:23, 6:45, 8:11, 8:12, 8:15, 8:28, 10:25, 12:28, 14:23, 14:26, 14:27, 14:28, 15:33, 16:31, 16:37, 16:51	23	10
Encryptie betekent het versleutelen van data, zodat alleen mensen die de ‘sleutel’ hebben de data kunnen lezen. Dat is in meer of mindere mate gebruikelijk	1:47, 2:37, 2:38, 2:39, 6:24, 6:25, 6:45, 6:46, 6:47, 9:34, 10:31, 15:34	12	6
Voorbeelden van quotes: 3:43: Interviewer: Veiligheid klinkt als een belangrijk onderdeel van de data hier. Wat doe je daar precies allemaal aan? Geïnterviewde: Vertrouwen op de case organisatie dat ze hun boeltje op orde hebben qua intruders. En ik heb gewoon enkel project shares met heel beperkte access lists dus ik weet exact wie bij welke data kan. 16:51: Geïnterviewde: Maar niet echt nagedacht over beveiliging en encryptie daarvan en het is altijd goed gegaan door de toegangssregels, wie er mag inloggen op de machine. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 72, Data beveiligen met authenticatie, autorisatie en encryptie

Voor sommigen is alleen authenticatie en autorisatie voldoende, ze hoeven niet terug te kunnen zien wie er allemaal toegang hebben gehad (logging) 3:44, 8:28.

Voor geïnterviewde 3 geldt dat deze met name de metadata versleutelt. De onderzoeksdata zijn niet te lezen zonder de metadata, dus dit is afdoende 3:45, 3:46, 3:47.

Twee onderzoekers zorgen dat gevoelige gegevens niet herleidbaar zijn naar een persoon 1:46, 5:49 (waardoor encryptie niet nodig is).

Twee onderzoekers geven aan geen encryptie te gebruiken 8:15, 12:26

11.7.2. Data beveiliging backup (Synchronisatie)

Data veiligstellen kan door data te synchroniseren[25].

Data beveiligen met backup en restore, bvlg backup			
Elf van de veertien geinterviewden geven aan dit te doen. Daar worden verschillende vormen voor toegepast	Genoemd in	#gen	#int
Automatische synchronisatie door een sync and share product als Dropbox of Surfdrive	1:29, 1:40, 3:46, 3:47, 12:18	5	3
Periodieke veelal handmatige kopieën naar andere opslag zoals een externe harddisk, een andere pc/laptop of een usb stick. Een bijzondere vorm hiervan is de data naar jezelf mailen	2:1, 2:55, 5:40, 5:41, 5:42, 6:20, 6:57, 8:21, 8:22, 8:23, 9:24, 12:16, 12:22, 14:10.	14	7
Men slaat de data op op centrale opslag en gaat uit van backup door de ICT afdeling	5:39, 5:46, 6:57, 14:10, 15:8, 16:20	6	5
Geen backup kan ook een optie zijn als de data eenvoudig opnieuw gegenereerd kunnen worden	9:3, 11:24, 14:10	3	3
Voorbeelden van quotes: 3:43: Interviewer: Veiligheid klinkt als een belangrijk onderdeel van de data hier. Wat doe je daar precies allemaal aan? Geïnterviewde: Vertrouwen op de case organisatie dat ze hun boeltje op orde hebben qua intruders. En ik heb gewoon enkel project shares met heel beperkte access lists dus ik weet exact wie bij welke data kan. 16:51: Geïnterviewde: Maar niet echt nagedacht over beveiliging en encryptie daarvan en het is altijd goed gegaan door de toegangssregels, wie er mag inloggen op de machine. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 73, Data beveiligen met backup en restore

Eén onderzoeker geeft aan dat hij ooit heeft moeten restoren (twee keer), hij gebruikte daar DropBox voor 1:29.

Vijf onderzoekers geven aan nooit een restore nodig gehad te hebben 2:85, 5:44, 6:22, 9:31, 12:20.

11.7.3. Data beveiliging contract

Soms zijn er contractuele verplichtingen om data extra te beveiligen (met autorisaties en/of encryptie) 8:19, 10:5, 11:34, 12:8, 14:35, 14:36, 14:37, 15:4 (acht quotes uit zes interviews).

[25] **Synchroniseren**: Data naar een tweede (of meer dan dat) locatie repliceren. Dat kan een eenmalige actie zijn (een kopie), het kan realtime zijn, waarbij de beide omgevingen voortdurend gelijk zijn (spiegelen) of dat data op gezette tijden wordt gerepliceerd.

11.7.4. Logging

Men is enigszins geïnteresseerd in logging om beveiligingsredenen. Men lijkt evenzeer geïnteresseerd in logging om te kunnen zien wat anderen aan activiteiten met hun data ontplooien.

Data beveiligen door logging			
Men kan om verschillende redenen geïnteresseerd zijn in logging	Genoemd in	#gen	#int
Geïnteresseerd in logging van gebruik door anderen (niet om beveiligingsredenen)	3:38, 3:39, 3:40, 8:17, 11:37, 16:52, 16:53, 16:54	8	4
Geïnteresseerd om beveiligingsredenen	1:33, 1:42, 8:39, 11:33, 15:35	5	4
Niet geïnteresseerd in logging om beveiliging	3:44, 5:53, 10:27, 14:24	4	4
Voorbeelden van quotes: 5:53: Interviewer: Is het voor jou van belang om te zien wie er allemaal aan de data heeft gezeten, dat je een soort van logging daarvan hebt? Geïnterviewde: Nee dat is niet nodig omdat ik niet met hele grote groepen mensen aan data analyse tegelijk werk 11:37: Interviewer: It's not something that's important for you? If it would not be there.... Geïnterviewde: It's not important, sometimes it's interesting to see that one paper is downloaded a lot. But that's not so important. What's more important is when they cite it, or use it, what they say about it. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 74, Data beveiligen door logging

Op basis van TABEL 74 zou voorzichtig geconcludeerd kunnen worden, dat de helft van de geïnterviewden geïnteresseerd is in logging. Deze groep is weer gelijk verdeeld over mensen die geïnteresseerd zijn om beveiligingsreden, om te weten hoe anderen hun data gebruiken of allebei.

Als men aangeeft niet geïnteresseerd te zijn in logging om veiligheidsredenen, betekent dat vaak dat het voldoende is, dat ze kunnen instellen wie er bij de data mag komen (authenticatie en autorisatie).

11.7.5. Co-Occurrence binnen data beveiligen categorie

De onderlinge verwantschappen binnen de beveiligingscategorie staan in TABEL 75.

Verwantschappen binnen de data beveiligen categorie			
	Code 1	Code 2	Coëfficiënt
1	Bvlg authenticatie	Bvlg autorisatie	0.47, 7
Uitleg: Er zit vooral veel verwantschap tussen authenticatie en autorisatie. Een systeem zal altijd de gebruiker verifiëren (authenticatie) voordat het toegang geeft tot data indien dat voor de gebruiker is toegestaan (autorisatie). Het lijkt er op dat het ook meespeelt, dat het onderscheid mogelijk niet geheel duidelijk is voor de geïnterviewden.			

Tabel 75, Verwantschappen binnen de data beveiligen categorie

11.7.6. Conclusie data beveiliging

Op basis van de interviews kan data beveiliging gedefinieerd worden als: Data beveiliging is het veilig stellen van de data en bestaat uit één of meer van de volgende aspecten:

- Autorisatie;
- Authenticatie;
- Encryptie;
- Backup (en restore);
- Contract (toegang alleen voor gecontracteerde derden).

Er zit een logische sterke verwantschap tussen Autorisatie en Authenticatie. Een systeem zal altijd de gebruiker verifiëren (authenticatie) voordat het toegang geeft tot data indien dat voor de gebruiker is toegestaan (autorisatie).

Het lijkt dat dat helft van de geïnterviewden geïnteresseerd is in logging. Deze groep is weer gelijk verdeeld over mensen die geïnteresseerd zijn om beveiligingsreden, om te weten hoe anderen hun data gebruiken of allebei.

11.8. Werkwijze

Hoe men inhoudelijk werkt als het o.a. gaat om naamconventies en folderstructuren en in mindere mate om datadocumentatie, versiebeheer en het gebruik van lab notebooks wordt in de respectievelijke hoofdstukken (

Metadata en DATADOCUMENTATIE) besproken. In de volgende paragraaf wordt meer naar de wijze waarop werk wordt afgestemd gekeken en in de daaropvolgende naar de motivatie.

Relevante codes in dit hoofdstuk staan in TABEL 76.

Subcodes voor de werkwijze	
Subcode omschrijving	Subcode in Atlas
Format gebruik folderstructuren in de groep	Md fdst groep
Format gebruik folderstucturen individueel	Md fdst persoonlijk
Format gebruik naamconventies groep	Md mc groep
Format gebruik naamconventies individueel	Md nmc persoonlijk
Werkwijze onderzoekers	Ind werkwijze
Gedrag onderzoeker flexibiliteit	Indmot flexibiliteit
Gedrag onderzoeker overhead	Indmot overhead
Integriteit	integriteit
Opmerkingen bij de codes De eerste vier gaan vooral over afstemming en keuzes in de onderzoeksgroep, kan men individueel invulling geven aan folderstructuren en naamconventie of is er een groepsformat? De codes over gedrag focussen zich op de manier waarop de onderzoeker werkt en kwamen tijdens de interviews naar voren. Het ging om keuzes die men maakte voor ondersteunende RDM tooling (vaak gebaseerd op pragmatiek), de al dan niet flexibele invulling van RDM (tooling) en hoe men omgaat met werkzaamheden die niet direct met het onderzoek te maken hebben. Integriteit kwam ook op tijdens de interviews en gaat vooral over de data die experimentele wetenschappers publiceren (alleen de geslaagde experimenten of ook de mislukte).	

Tabel 76, Subcodes voor de werkwijze

11.8.1. Wijze waarop werk wordt afgestemd

Als eerste het niveau[26] waarop afstemming over de invulling van de werkzaamheden wordt besproken.

Dit zijn de niveaus zoals ze in de universiteit bestaan:

1. Afdelingsniveau;
2. Sectieniveau;
3. Onderzoeksgroepniveau;
4. Individueel niveau.

In de interviews worden geen voorbeelden gevonden van afspraken over invulling van werkzaamheden op afdelings- of sectieniveau, wel op groepsniveau. Een groep of individu kan een verplichte werkwijze krijgen (protocollen van de groep bijvoorbeeld), een geadviseerde werkwijze (senior onderzoekers die de 'best practices' aangeven) en een vrije werkwijze (waarin individuen in een onderzoek hun eigen keuzes maken). Dit staat samengevat in onderstaande TABEL 77.

[26] Faculteiten worden opgedeeld in afdelingen, de meeste afdelingen hebben secties. Onderzoeksgroepen zitten typisch in een sectie, maar kunnen ook bestaan uit samenwerkingsverbanden van wetenschappers (uit andere afdelingen) binnen of buiten de case organisatie.

mate van vrijheid in werkwijze afgezet tegen individu en groep, Md fdst groep, Md fdst persoonlijk, Md mc groep, Md nmc persoonlijk			
	opgelegd	Geadviseerd	vrij
Groepsniveau	1:36, 1:37, 8:11, 831, 10:16	1:36, 1:37, 2:25, 3:42, 3:62; 7:65, 10:14	
Individueel niveau		3:42, 3:62, 7:65, 10:14, 16:24	2:16, 2:42, 5:58, 5:59, 6:28, 9:13, 9:16, 11:45, 14:17, 16:3
Opmerkingen: Een onderzoeker kan zijn/haar eigen keuzes maken over de invulling van de werkzaamheden, krijgt hier advies over van de onderzoeksleider of volgt het protocol van de groep. Het advies (zoals benoemd in de interviews) is nooit bindend. Mensen kunnen meestal alsnog zelf kiezen hoe ze invulling geven. 1:36 en 1:37 vallen er een beetje buiten. De onderzoeker stelt dat hij een soort beleid opstelt, maar dat hij niet over de schouders mee gaat kijken of het ook gevolgd wordt. Het is niet 100% opgelegd en het is niet 100% vrij. Onderzoeker 10 legt de folderstructuur op en geeft een advies over de naamconventies. 8:11 en 8:31 vallen helemaal buiten de gewoontes van de andere groepen. De begeleiders slaan hier alles zelf op in een vastliggende structuur (individuele onderzoekers kunnen dat dus niet zelf). Voorbeelden van quotes: 5:53: Interviewer: Is het voor jou van belang om te zien wie er allemaal aan de data heeft gezeten, dat je een soort van logging daarvan hebt? Geïnterviewde: Nee dat is niet nodig omdat ik niet met hele grote groepen mensen aan data analyse tegelijk werk 11:37: Interviewer: It's not something that's important for you? If it would not be there.... Geïnterviewde: It's not important, sometimes it's interesting to see that one paper is downloaded a lot. But that's not so important. What's more important is when they cite it, or use it, what they say about it. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 77, mate van vrijheid in werkwijze afgezet tegen individu en groep

Het opleggen van hoe men individueel moet werken (met name folderstructuur en naamconventie) wordt in de interviews niet benoemd. Het opleggen van een folderstructuur op groepsniveau wel. Groepsleiders die een advies geven, de één wat dwingender dan de ander, komt vaker voor in de interviews. Als mensen vooral zelfstandig werken, lijken ze de invulling van hun werkzaamheden helemaal zelf te bepalen. Het zwaartepunt in de tabel lijkt meer in de richting van vrij dan in de richting van opgelegd te gaan.

Verschillende artikelen uit het literatuuronderzoek geven ook een indruk over de omgeving (universiteiten) waarin de RDM gerichte onderzoeken hebben plaatsgevonden.

- Een universiteit is een gedecentraliseerde organisatie, waarin veel verschillende culturen huizen (Cox et al. 2016; Flores et al. 2015).
- Faculteiten, afdelingen, onderzoeksprojecten en onderzoekers hebben vaak hun eigen gewoonten als het gaat om onderzoek doen en ook hun eigen ideeën als het gaat om wat goede RDM praktijken zijn (Burgi et al. 2017; Cox et al. 2016).
- Inspanningen om een gecentraliseerd RDM programma op te zetten kunnen gezien worden als een poging om deze verschillende culturen met hun gebruiken samen te brengen (Wittenberg and Elings 2017).
- RDM kan echter alleen effectief zijn als het aansluit op de praktijk en cultuur van de individuele onderzoeker met zijn eigen instrumenten, onderzoekstools en IT omgeving (Jones 2013).

De stellingen uit de literatuur en de uitkomsten van de interviews lijken goed in lijn te liggen, met die beperking op de stellingen uit de literatuur dat onderzoeksgewoonten worden bepaald op het niveau van onderzoeksgroepen en/of individuen.

- Onderzoeksgewoonten worden op onderzoeksgroep- en veelal op individueel niveau bepaald, men heeft daarin ogenschijnlijk veel vrijheid in de keuze van werkwijze, zeker als men individueel onderzoek doet.
- Voor een RDM oplossing voor opslag geldt, dat de verschillende onderzoeksgewoonten hier op moeten kunnen aansluiten[27].

11.8.2. Individuele motivatie

In de interviews komen verschillende individuele beweegredenen voor bepaalde werkwijzes of het gebruik van bepaalde tools aan de orde. De wetenschappers zijn vaak zeer doelgericht met hun onderzoek bezig. Alles wat niet met het wetenschappelijke probleem te maken heeft, wordt meestal gezien als overhead. Als het specifiek gaat om RDM tooling, dan is er een verschil in de mate van flexibiliteit en automatisering die men zoekt. Het wordt hieronder één voor één behandeld:

11.8.3. Individuele motivatie flexibiliteit

Elk project is anders. Flexibiliteit en functionaliteit gaan daarom altijd voor een vaste structuur waarin elk project moet passen (3:42, 3:59, 3:60, 3:62, 3:63, 9:13, 9:16, 9:29, 14:40) en: Hoe iemand zijn project structureert is een individuele keuze (11:45, 14:17) (elf quotes uit vier interviews).

Voorbeeld quotes:
3:42: Geïnterviewde: Het belangrijkste is dat het Free form is, dat je er mee kan doen wat je wil dat het niet beperkend werkt.
3:62: Geïnterviewde: Flexibiliteit is altijd belangrijker dan alle regeltjes volgen vandaar dat het eerder een suggestie is dan een fixed set of rules en er zijn wel een aantal suggesties zo van: doe het op deze manier, daar ga je me binnen twee jaar dankbaar voor zijn.

11.8.4. Individuele motivatie overhead

Tools die het onderzoek ondersteunen dienen een zo laag mogelijke leercurve te hebben en moeten meteen, zonder inspanning, werken. Anders gezegd: Het moet zo min mogelijk tijd en moeite kosten om een ondersteunende RDM tool voor het onderzoek te kunnen gebruiken. (1:60, 1:61, 5:33, 5:34, 6:42, 6:44, 7:7, 9:17, 11:20) (negen quotes uit zes interviews).

Als de tools daar niet aan voldoen, kan het ook een barrière zijn in je wetenschappelijke werkzaamheden. (5:35, 10:24). Een genoemd voorbeeld is dat mensen dan minder geneigd zijn te delen (5:35).

Afwijken van wat je al jaren gebruikt is sowieso lastig, men moet aan nieuwe manieren wennen (1:59, 6:44, 7:6, 7:7).

Voorbeeld quotes:
9:17: Interviewer: Dat heb ik ook meer gemerkt[in eerdere interviews]. Het is toch de tools gebruiken met een zo laag mogelijke leercurve, het moet gewoon werken. Geïnterviewde: Ja precies want de tools zijn ook dermate ingewikkeld dat ik er liever minder aandacht besteed aan de tools dan hoe ga ik de data verwerken. Want dat is het wetenschappelijke probleem ook niet.
5:33: Geïnterviewde: Ja, Zo gemakkelijk mogelijk en ook zo snel mogelijk. Ja het is toch de werkdruk die veel wetenschappers hebben om zo snel mogelijk dingen te doen. Er zijn heel veel dingen die

moeten gebeuren. (5:34) Interviewer: Alles wat ondersteunend is, moet een zo laag mogelijke leercurve hebben en het moet werken? Geïnterviewde: Ja precies.

Het moet zo min mogelijk tijd en moeite kosten om een ondersteunende RDM tool voor het onderzoek te kunnen gebruiken. Vaak betekent dat dat men blijft bij de tools die men al kent.

11.8.5. Integriteit

Integer omgaan met data wordt op twee manieren benoemd in de interviews (TABEL 78).

Werkwijze integriteit, integriteit			
Men kan om verschillende redenen geïnteresseerd zijn in logging	Genoemd in	#gen	#int
Bij experimentele wetenschap komen vaak alleen links naar de datasets van de geslaagde experimenten in de publicatie. De datasets van de niet geslaagde experimenten worden vaak niet opgeslagen en/of benoemd	1:7, 1:8, 2:52, 2:73, 2:74, 2:75, 7:66	7	3
Voor peer review is het noodzakelijk om het onderzoek te kunnen reproduceren. Dit bepaalt welke data er beschikbaar gesteld moet worden aan de peer reviewer en dus ook wat er uiteindelijk in het archief moet om aan de policy van de case organisatie te voldoen (Dunning et al. 2018) en aan die van sommige journals	5:13, 5:25, 5:26, 5:27, 9:5, 9:6, 9:7	7	2
Voorbeelden van quotes: 2:75: Geïnterviewde: People don't say what didn't work people only report what did work and that's quite tricky. I mean now of course is the discussion of research integrity and I'm not blaming that people are doing bad science. There are a lot of people who are trying to do the good thing but if you publish what didn't work that's not interesting for publishers and as a result publications are always written in a way as if we first did this and we got this very nice result and now we did that. But people don't explain what's happened in between. And that is quite upsetting because it's really fake. I mean fake in the sense that we cut out the parts which didn't work out 9:6: Interviewer: Interviewer: En dat is om de validiteit van het onderzoek te kunnen onderzoeken? Geïnterviewde: Ja want het tijdschrift vereist gewoon dat als je iets publiceert dan moet iemand anders met alle gegevens die in het artikel staan kunnen reproduceren wat jij geproduceerd hebt. Interviewer: Is dat dan zo simpel van: Hij heeft de publicatie, hij heeft de dataset en dan moet hij het na kunnen spelen? Geïnterviewde: In feite wel. Dus gegeven dat de algoritmes gewoon open source zijn en gegeven dat die bekend zijn en gegeven hoe het moleculaire model eruitziet moet je gewoon met alle gegevens opnieuw de data kunnen regenereren. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 78, Werkwijze integriteit

Integriteit lijkt er vooral op gericht te zijn om zo open mogelijk te laten zien hoe het onderzoek is verricht, welke stappen men heeft gezet (zowel de geslaagde als de mislukte), welke keuzes men heeft gemaakt en waarom.

11.8.6. Co-Occurrence bij werkwijze

Behalve de al benoemde code co-occurrence van de md codes in het metadata hoofdstuk (zie

CODE CO-OCCURRENCE BINNEN DE METADATA CATEGORIE) zijn er geen andere voldoende sterke verwantschappen gevonden.

11.8.7. Werkwijze conclusie

Als wordt gekeken naar de werkwijze van de onderzoekers, kan men concluderen dat:

- Onderzoeksgewoonten op de case organisatie worden op onderzoeksgroep- en veelal op individueel niveau bepaald, men heeft daarin ogenschijnlijk veel vrijheid in de keuze van werkwijze, zeker als men individueel onderzoek doet;
- Voor een RDM oplossing geldt, dat de verschillende onderzoeksgewoonten hier op moeten kunnen aansluiten.

Generieke onderzoeksgewoonten zijn:

- Een wetenschapper wil zich vooral bezighouden met het doen van onderzoek. Allerlei ondersteunende werkzaamheden moeten zodanig tot het minimum beperkt worden, dat ze voldoende zijn voor het eindresultaat, maar niet zo veel tijd kosten dat er minder tijd overblijft voor het onderzoek;
- Het moet zo min mogelijk tijd en moeite kosten om een ondersteunende RDM tool voor het onderzoek te kunnen gebruiken. De consequentie hiervan is vaak dat men blijft bij de tools die men al kent;
- Elk project is anders. Flexibiliteit en functionaliteit gaan daarom altijd voor een vaste structuur waarin elk project moet passen. RDM tooling moet deze flexibiliteit en functionaliteit kunnen bieden.

T.a.v. integriteit kan gezegd worden:

Integriteit lijkt er vooral op gericht te zijn om zo open mogelijk te laten zien hoe het onderzoek is verricht, welke stappen men heeft gezet (zowel de geslaagde als de mislukte), welke keuzes men heeft gemaakt en waarom.

11.9. Vindbaarheid

Vindbaarheid (voor dit onderzoek): De stakeholders in de managing active data fase van een onderzoek weten waar de data van een onderzoek staat. De bijbehorende vraag is: ”Waar zijn de data (fysiek en/of logisch)?

Uit de interviews blijkt dat het probleem dat data verspreid over diverse opslagmedia staan herkenbaar is (zie ook DATA OPSLAG). Codes die voor de vindbaarheid van de data naar voren komen staan in TABEL 79.

Subcodes in de codecategorie vindbaarheid	
Subcode omschrijving	Subcode in Atlas
Folderstructuren als hulpmiddel om data te vinden	vind folderstructuur
Naamconventies als hulpmiddel om data te vinden	vind naamconventie
ELN’s als hulpmiddel om data te vinden	vind ELN
In staat zijn om succesvol data terug te vinden	vind ja

Tabel 79, Subcodes in de codecategorie vindbaarheid

11.9.1. Vindbaarheid van data

Manieren die in de interviews genoemd worden om data vindbaar te houden, zijn o.a. folderstructuren, naamconventies en ELN’s (TABEL 80).

Vindbaarheid met behulp van folderstucturen, naamconventies en ELN’s, vind folderstructuur, vind naamconventie, vind ELN			
Men kan om verschillende redenen geïnteresseerd zijn in logging	Genoemd in	#gen	#int
Door een vaste folderstructuur te gebruiken, weet men welk type data (research data, metadata, datadocumentatie, artikelen, enz.) men waar kan vinden.	1:35, 1:52, 2:41, 5:57, 5:64, 6:27, 6:48, 7:62, 8:7, 8:8, 8:10, 8:31, 9:22, 10:18, 11:38, 11:40, 12:35, 16:56, 16:59	19	11
Door specifieke naamgeving te gebruiken, kan men de inhoud van een dataset herkennen aan de naam en zo vinden.	2:64, 5:56, 7:64, 8:5, 8:9, 12:4, 16:59, 16:63	8	6
Soms wordt in de ELN’s bijgehouden waar data staan opgeslagen of vindt men dat een goed idee, het voordeel is dat deze logboeken doorzoekbaar zijn.	1:55, 3:40, 7:58, 7:59, 7:60, 10:19, 12:11, 12:12, 12:35	9	5
Voorbeelden van quotes: 5:64: Interviewer: Maar sla je dat weer zo op, dat je het makkelijk terug kan vinden? Geïnterviewde: Dat zit gewoon in die mappen. Het probleem is alleen dat ik soms te veel opsla en dat ik niet meer weet waar ik het moet zoeken. Dat je zoveel bestanden hebt dat je denkt, waar stond het nou ook alweer? 16:56: Interviewer: Maar jij hebt ook niet voor jezelf een eigen conventie, dat is per project anders? Geïnterviewde: Nee, behalve de gewone dingen zoals Temp is voor tijdelijk. Wat je ook vaak wel ziet is dat de data opgesplitst is in geografische gebieden en dat we coördinaten gebruiken om delen van data in verschillende directories te stoppen. Maar dat is eigenlijk wel per project anders, dat is de realiteit. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 80, Vindbaarheid met behulp van folderstucturen, naamconventies en ELN’s

11.9.2. Vindbaarheid, kunnen vinden

Wetenschappers kunnen hun data vinden door folderstructuren, naamconventies en verwijzingen in lablogboeken te gebruiken. Soms aangevuld met zoekfunctionaliteit (1:35, 1:41, 1:52, 1:55, 2:64, 3:40, 5:64, 6:27, 7:58, 7:59, 7:60, 7:62, 7:64, 8:5, 8:7, 8:8, 8:9, 8:10, 8:31, 9:22, 9:23, 10:19, 11:15,

11:38, 11:40, 11:47, 12:11, 12:12, 12:35, 16:37, 16:56, 16:58, 16:59, 16:63) (code 'vind ja' 34 keer genoemd in twaalf interviews).

11.9.3. Vindbaarheid definitie

Voorafgaand is de conclusie al getrokken dat actieve fase niet meer gebruikt wordt. Bovendien is vindbaarheid toegespitst geweest op de onderzoekers. Hoewel er twee data stewards en een IT support medewerker in een afdeling zijn geïnterviewd, spraken ook zij meer over hoe onderzoekers in dit verband werken, dan hoe zijzelf of andere stakeholders dat doen. De definitie wordt daarom veranderd naar:

Vindbaarheid (voor dit onderzoek): De onderzoekers in een lopend onderzoek weten waar de data van een onderzoek staan. De bijbehorende vraag is: "Waar zijn de data (fysiek en/of logisch)?

11.9.4. Co-occurrence binnen de vindbaarheid categorie

In TABEL 81 staan de verwantschappen tussen de verschillende codes beschreven.

Verwantschappen binnen de vindbaarheid categorie			
	Code 1	Code 2	Coëfficiënt
1	vind ja	vind folderstructuur	0.38, 8
2	vind ja	vind notebook	0.26, 5

Uitleg:
De waardes in de tabel suggereren met name dat de vindbaarheid van data wordt bevorderd door:

- Het gebruik maken van een folderstructuur;
- Het gebruik maken van (verwijzingen in) lablogboeken.

Tabel 81, Verwantschappen binnen de vindbaarheid categorie

11.9.5. Vindbaarheid conclusie

Voor vindbaarheid blijft de definitie uit het literatuuronderzoek staan. Voor vindbaarheid geldt dat de customers en participants van het werksysteem in staat zijn om de relevante data van het lopende onderzoek te vinden. Het verband tussen folderstructuren en vindbaarheid is sterker dan voor volgbaarheid. Door per onderzoek een bekende folderstructuur te gebruiken en die consequent te hanteren, kan men relevante data eenvoudig terugvinden. Een andere sterke verwantschap is het verband tussen een ELN en vindbaarheid: In een ELN kan men een verwijzing maken (bijvoorbeeld een snelkoppeling) naar een dataset. Dat maakt terugvinden eenvoudiger. Door de namen van de databestanden, die bij een onderzoeksstap horen, te voorzien van bepaalde beschrijvende parameters (denk aan een snelheid in het experiment, of de datum dat de onderzoeksstap werd uitgevoerd) kan daar op gezocht worden (vindbaarheid) maar het kan ook helpen met volgbaarheid (als er elke datum een andere snelheid wordt gebruikt bijvoorbeeld). (zie ook CONCLUSIE VERWANTSCHAPPEN CROSS-CATEGORIE)

Data worden opgeslagen op diverse centrale systemen (door de case organisatie aangeboden), op Share and Sync oplossingen op externe datahouders en code data op systemen die daar specifiek voor bedoeld zijn.

11.10. Volgbaarheid

Volgbaarheid (voor dit onderzoek): Data kunnen in elke fase van het onderzoek geïnterpreteerd worden en de tussentijds toegepaste transformaties zijn gedocumenteerd. De bijbehorende vraag is: ”Kan ik de wijzigingen die de data ondergaan volgen op basis van de beschikbare informatie?” Het gaat hierbij om de gewijzigde data zelf, de informatie over de wijzigingen en de hulpmiddelen die helpen de wijzigingen te herkennen.

In één van de interviews wordt de term data provenance genoemd (10:2, 10:3). Op de RDM website van de VU staat hierover (vertaald uit het Engels):
“De term "data provenance" verwijst naar noodzakelijke registraties die de oorsprong van een gegeven (in een database, document of opslagplaats) verklaren, samen met een uitleg over hoe en waarom het op de huidige plaats terecht is gekomen (Encyclopedia of Database Systems, pp 608-608). Je kunt het ook het proces van het bijhouden van wijzigingen in de gegevens noemen.” (Anon 2020)

Er wordt bovendien verwezen naar labboeken en dagboeken die hiervoor gebruikt kunnen worden.

De **definitie** van volgbaarheid kan daarmee veranderd worden naar:
Volgbaarheid omvat de data provenance, het proces van het bijhouden van wijzigingen in de onderzoeksdata, inclusief de middelen waarin die wijzigingen worden bijgehouden.

In de interviews komen de subcodes zoals vermeld in naar voren (TABEL 82).

Subcodes in de codecategorie volgbaarheid	
Subcode omschrijving	Subcode in Atlas
Het benoemen en kunnen volgen van de verschillende versies (versiebeheer)	volg versie
Volgen met behulp van naamconventies	volg naamconventie
Volgen met behulp van folderstructuren	volg folderstructuren
Het kunnen volgen van aanpassingen aan code	volg code
Het bijhouden van de transformaties	volg transformatie
Het in staat zijn om succesvol te volgen	volg ja

Tabel 82, Subcodes in de codecategorie volgbaarheid

11.10.1. Volgbaarheid versies

Versies van de data worden op verschillende manieren bijgehouden. Over het algemeen gaat het in deze code vooral over het bijhouden van het feit dat er wijzigingen zijn, het gaat meestal niet over hetgeen daadwerkelijk gewijzigd is. De uitzondering is geïnterviewde 8, die in de metadata voortdurend de toestand van zijn sensoren bijhoudt en met die informatie kan achterhalen waarom plotselinge veranderingen plaatsvinden. Dit staat beschreven in TABEL 83.

Volgbaarheid met versiebeheer, volg versie			
Methoden van versieregistratie	Genoemd in	#gen	#int
Met versiebeheer van de Sync and Share software (Dropbox)	1:48, 2:11, 2:70	3	2
Door een volgnummer toe te voegen aan de bestandsnaam van de data	2:66, 5:55, 5:56, 5:61, 6:29, 16:61, 16:63	7	4
In de metadata	1:51, 8:3, 8:16, 8:17, 12:6	5	3
Door de ingebouwde versiebeheer functionaliteit van GitHub, GitLab en/of Bit Bucket te gebruiken (vooral voor code)	6:32, 6:33, 6:50, 9:11, 11:10, 11:14, 11:18, 14:20, 15:44	9	5
Voorbeelden van quotes: 1:48: Interviewer: Do you use some sort of versioning on your data sets? Well my imagination is it like you get your raw data and you do some manipulation and you get a resulting set with version 2 or something and then you know maybe more manipulation version 3 and you record whatever you did between the versions…. Geïnterviewde: I guess try to but the problem is I don't really do a lot of the research myself so it's best as I can try to get the students to do it. Naming conventions I definitely have. So I actually have a little naming file, naming conventions file that I share for every students they have project names and dates for naming but for documents I do have version control. So there again I have a Dropbox. 16:61: Interviewer: Maar voor data doe je dat niet zodat je bijvoorbeeld de eerste dataset een v1 geeft en na manipulatie een v2 en dat je tussentijds bijhoudt wat er is gebeurt? Geïnterviewde: Niet systematisch nee, er staat misschien, eh, als er al conventie is, is het dat er een source achter staat, dat het de brondata is, of de definitieve dataset die heeft dan de naam van die dataset wellicht. Final, maar dat durf ik niet eens met zekerheid te zeggen. Of dat we misschien een tussen versie 1 hebben. Wat ik zelf wel doe met al mijn documenten, maar dat vind ik dan minder data, ik gebruik extreem systematisch V1 en V2 V3 V4. Soms heb ik wel 100 versies staan. En dan denk ik: Ok is wel een beetje veel. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 83, Volgbaarheid met versiebeheer

11.10.2. Volgbaarheid naamconventie

De tweede bullet van de vorige paragraaf geeft al een aanwijzing dat de naamconventie een rol kan spelen bij het bijhouden van wijzigingen. Soms beschrijft de naamgeving de bewerkingsstap die op de data is uitgevoerd, tezamen met een volgnummer of een datum, zou je hieruit summier de wijziging uit kunnen afleiden. De mate van detail wisselt per onderzoeker.

- Naamconventie die de dataset beschrijft (als men twee opeenvolgende datasets heeft, kan men een wijziging afleiden (1:49, 2:64, 5:56, 8:5; 10:14, 10:17, 11:40, 12:4, 12:6, 12:35, 16:58, 16:61) (code 'volg naamconventie' twaalf quotes in acht interviews).

Het vinden van dat een wijziging heeft plaatsgevonden (waar mogelijk een versie een indicatie van is) en wanneer, vergt dus wat uitzoekwerk. Wat de wijziging precies is geweest, staat is hooguit beperkt terug te halen uit de naamconventie als men in de naam beschrijft wat er in de onderzoeksstap heeft plaatsgevonden (2:64, 8:5, 10:14, 11:40, 12:4). Het zou voor de hand liggen dat meer details zijn terug te vinden in datadocumentatie en/of metadata.

Voorbeelden van quotes zijn:

10:14: Geïnterviewde: En heb je een eigen naamgeving? Of is dat iets wat jullie met de afdeling gebruiken? Geïnterviewde: Ik heb mijn eigen conventie maar dat is dan bijvoorbeeld ok: n experiments x strategies zoiets weet je wel, dat soort dingen.....

12:4: Geïnterviewde: And what's nice is that you can enter some meta data there. And what we tend to do it's become a kind of within the research group. We tend to automatically put a fairly descriptive file name on it. It's the kind of thing that you think one person starts doing and it's kind

of spread. So automatic so you can set up a kind of template for the file name and then you just change your put in the individual. I can show it to you if you want to later.

11.10.3. Volgbaarheid folderstructuur

De folderstructuur werkt altijd in combinatie met de naamconventie als het om volgbaarheid gaat. Geïnterviewde 11 geeft een datum mee aan de data analyses die worden gemaakt en bewaart ze in een daarvoor bestemde folder. Men kan daarmee de stappen van de databewerking afleiden (11:38, 11:40, 11:42).

Geïnterviewde 12 gebruikt folders met een datum in het ELN. De datasets krijgen een behoorlijke beschrijvende naam van de inhoud ('fairly descriptive file name' 12:4). Op basis hiervan zijn de wijzigingen weer te herleiden (12:35).

Het vinden dat een wijziging heeft plaatsgevonden (waar mogelijk een versie een indicatie van is) en wanneer, vergt dus wat uitzoekwerk. Wat de wijziging precies is geweest, staat is niet terug te halen uit de folderstructuren. Het zou voor de hand liggen dat dit wel terug te vinden is in datadocumentatie en/of metadata.

11.10.4. Volgbaarheid code

De aanpassingen die aan code worden gedaan, zijn vaak wat beter te volgen aangezien men daar over het algemeen gespecialiseerde applicaties voor gebruikt (meest genoemd Github, Gitlab en Bit bucket) (1:27, 6:32, 6:33, 8:25, 9:11, 10:17, 11:10, 11:14, 14:18, 14:20, 15:44, 16:60) (code 'vlg code' twaalf quotes in negen interviews).

Sommigen gebruiken dit ook voor hun documentatie zoals in de quotes wordt benoemd.

Voorbeelden van quotes:
6:32: Interviewer: En data documentatie? Geïnterviewde: Dat gebeurt, ehh, ligt er ook heel erg aan natuurlijk want sommigen zitten volledig geautomatiseerd en doen alles met Git en dat gaat als een trein.
14:20: Interviewer: en gebruik je dat ook voor versioning van zowel de code als van de papers? Geïnterviewde: Ja precies dus als er met meerdere mensen wordt gewerkt, wordt het via Bit Bucket gedaan en de data worden ook gelijk neergezet bij elkaar en weet je precies hoe het is gedaan als er een revisie moet komen. Kunnen we dat makkelijk doen.

11.10.5. Volgbaarheid Transformaties

In deze code gaat het om inhoudelijke informatie over de transformatie (TABEL 84).

Volgbaarheid van transformaties, vlg transformatie			
Volgen van transformaties met	Genoemd in	#gen	#int
Naamconventies	1:48, 1:49, 5:55, 12:4, 16:61, 16:63	6	4
Datadocumentatie in een ELN	1:55, 1:56, 3:37[28], 3:38, 3:39, 3:40; 10:16, 15:42	8	4
Datadocumentatie in een los bestand	5:61, 12:39, 15:42	3	3
Door de ingebouwde versiebeheer functionaliteit van GitHub, GitLab en/of Bit Bucket te gebruiken (vooral voor code)	8:25, 9:11, 10:17, 11:10, 11:14, 11:42	6	4
Met behulp van werkafspraken, men spreekt af dat dat in een bepaalde folder door niemand mag worden gewijzigd	3:34, 3:35, 11:38, 11:40, 11:42[29]	5	2
De metadata	1:51, 8:3, 8:16, 8:17, 11:42	4	2
Voorbeelden van quotes: 8:25: Interviewer: And do you store some kind of information about a version, like this version I did this manipulation for it to get to this version and I did use this manipulation to get to the next version. Geïnterviewde 2: Yeah. in coding. Yes. I use indicators to show the change and what will this change entails. Yes. 11:42: Geïnterviewde: And so I mean I run my code and basically read some input file, but this input file only contains things that you want to change. And there is an outputfolder that includes the Git commit number. And then all the possible settings that you could have changed. But then the default. So that this file could be also used as an input file again to exactly reproduce that so that. So that's basically all a meta data. (alle benoemde quotes zijn terug te vinden in de BIJLAGE QUOTES UIT DE INTERVIEWS).			

Tabel 84, Volgbaarheid van transformaties

11.10.6. Volgbaarheid, kunnen volgen

In deze code ('volg ja') geven de geïnterviewden aan dat ze vinden dat ze de wijzigingen in de data registreren en soms ook wat de inhoud van de wijziging is. Voor de citaten die met deze code zijn gemarkeerd geldt dat ze minimaal met één van de voorgaande codes zijn gemarkeerd (1:27, 1:30, 1:48, 1:49, 1:51, 1:52, 1:55, 1:56, 2:11, 2:49, 2:51, 2:64, 2:69, 3:34, 3:35, 3:37, 3:38, 3:39, 3:40, 5:50, 5:55, 6:32, 6:33, 8:3, 8:4, 8:5, 8:16, 8:17, 8:18, 8:25, 9:11, 9:15, 10:13, 10:16, 10:17, 11:10, 11:14, 11:38, 11:40, 11:42, 11:47, 12:4, 12:5, 12:6, 12:35, 12:39, 14:20, 15:42, 15:44, 16:60) (50 keer genoemd in veertien interviews, voorbeelden van quotes staan in de eerdere paragrafen).

[28] Bij geïnterviewde 3 gaat het wegschrijven naar het labboek automatisch.

[29] Bij geïnterviewde 3 worden sommige datafolders read only gemaakt, niemand kan ze dan wijzigen, bij geïnterviewde 11 heeft men afgesproken dat data in bepaalde folders niet gewijzigd mogen worden.

11.10.7. Co-occurrence binnen de volgbaarheid categorie

In TABEL 85 wordt de verwantschap getoond tussen de verschillende codes van volgbaarheid.

Verwantschappen binnen de volgbaarheid categorie			
	Code 1	Code 2	Coëfficiënt
1	volg ja	Volg transformatie	0.49, 9
2	Volg versie	Volg code	0.31, 6
3	Volg ja	volg code	0.22, 9
4	Volg naamconventie	Volg folderstructuur	0.21, 2
5	Volg ja	Volg versie	0.20, 9
6	Volg transformatie	Volg versie	0.20, 5

Uitleg:

1. Er zijn twee sterke verwantschappen in deze codereeks. Met name tussen kunnen volgen en transformaties. Dat ligt voor de hand, als je weet wat de transformaties zijn, kun je per definitie volgen.
2. Een andere sterke verwantschap ligt tussen code en versiebeheer. Deze ligt voor de hand omdat voor code veelal geautomatiseerde version control systemen worden gebruikt (Github, Gitlab en Bit Bucket) die dit proces automatiseren. Mogelijk helpt het dat versiebeheer automatisch gaat (zie ook Individuele motivatie overhead).
3. De verklaring onder punt twee geldt ook voor het succesvol kunnen volgen van de data als het om code gaat. De waarde is wat lager omdat lang niet alle geïnterviewden met code werken, maar wel allemaal zijn bevraagd op volgbaarheid.
4. Eerder is ook al zichtbaar geworden dat mensen vaak en folderstructuren en naamconventies gebruiken. Deze midddelen worden ook ingezet voor de volgbaarheid. Het verband hier levert geen extra input op als het gaat om beter of minder goed kunnen volgen.
5. Goed versiebeheer bevordert wellicht het de volgbaarheid van de data. Het kan zijn dat veel mensen andere middelen gebruiken om data volgbaar te maken wat een wat zwakkere coëfficiënt verklaart.
6. Een analoge verklaring aan punt 5 geldt voor punt 6.

Tabel 85, Verwantschappen binnen de volgbaarheid categorie

Versiebeheer op met name code wordt vaak toegepast, waarschijnlijk omdat de systemen die gebruikt worden voor de opslag van code dit automatisch bieden. Dit leidt vaker dan bij andere data tot succesvolle volgbaarheid. Dit sluit aan bij de eerdere constatering dat wetenschappers graag gebruik maken van ondersteunende tooling die ze op relatief eenvoudige wijze, geautomatiseerd, werk uit handen neemt dat mogelijk niet direct met de essentie van hun onderzoek te maken heeft (INDIVIDUELE MOTIVATIE en INDIVIDUELE MOTIVATIE OVERHEAD). Versiebeheer lijkt in TABEL 85 het krachtigste middel te zijn om volgbaarheid te ondersteunen.

11.10.8. Volgbaarheid conclusie

Volgbaarheid omvat de data provenance, het proces van het bijhouden van wijzigingen in de onderzoeksdata, inclusief de middelen waarin die wijzigingen worden bijgehouden.

De invloed van versiebeheer, naamconventies en folderstructuren op de volgbaarheid van data is aanwezig, maar zwak. Een uitzondering is versiebeheer op code via daarvoor specifiek bedoelde systemen. Deze werkwijze vertoont ook een hoge mate van volgbaarheid. Goede beschrijvingen van onderzoeksdata en de vastlegging van hoe data zijn verzameld en achtereenvolgens zijn bewerkt in een ELN, zijn tezamen de sterkste vorm van volgbaarheid die is gevonden in dit onderzoek. De data volledig kunnen volgen, gebeurt allen bij geïnterviewde 3 en 8. Geïnterviewde 1 geeft aan hier interesse in te hebben. De anderen lijken daar minder in geïnteresseerd (zie ook CONCLUSIE VERWANTSCHAPPEN CROSS-CATEGORIE).

Versiebeheer op code wordt vaak toegepast, waarschijnlijk omdat de systemen die gebruikt worden voor de opslag van code dit automatisch bieden. Dit leidt vaker dan bij andere data tot succesvolle

volgbaarheid. Dit succes is mogelijk te danken aan het feit dat de deze systemen op relatief eenvoudige wijze, geautomatiseerd werk uit handen nemen van de wetenschappers.

11.11. Archiveerbaarheid

Archiveerbaarheid (voor dit onderzoek): Het mogelijk maken om alle voor het onderzoek relevante data (dit ter beoordeling aan de onderzoeker) te verzamelen en klaar te maken voor archivering. De bijbehorende vraag is: ”Wat zijn de relevante data die gearchiveerd moeten worden en hoe kan ik ze als dataset herkennen?”

Archiveren is iets wat je doet als de actieve fase van het onderzoek is afgelopen. Tijdens het onderzoek moet je echter gaan bepalen wat gearchiveerd moet worden en dan met name welke data en informatie nodig zijn om het onderzoek te kunnen valideren via de peer review (5:25, 5:26 en 9:5, 9:6) en om te voldoen aan de TU policy (2:78).

In de interviews is t.a.v. archiveerbaarheid naar voren gekomen wat er wordt gearchiveerd, waar, hoe en aan welke eisen een archief moet voldoen TABEL 86. Uiteraard is er ook gekeken of mensen überhaupt archiveren.

Subcodes in de codecategorie archiveerbaarheid	
Subcode omschrijving	Subcode in Atlas
Het benoemen en kunnen volgen van de verschillende versies (versiebeheer)	volg versie
Volgen met behulp van naamconventies	volg naamconventie
Volgen met behulp van folderstructuren	volg folderstructuren
Het kunnen volgen van aanpassingen aan code	volg code
Het kunnen volgen als er gedeeld wordt	volg delen
Het bijhouden van de transformaties	volg transformatie
Het in staat zijn om succesvol te volgen	volg ja

Tabel 86, Subcodes in de codecategorie archiveerbaarheid

11.11.1. Archiveerbaarheid wat

Geïnterviewde één vraagt zich af wat er gearchiveerd moet worden. Een overweging die vergelijkbaar is met die van geïnterviewde zeven over ruwe data en niveaus van data (7:34, 7:35, 7:36, 7:37). Alleen het tabelletje waarmee een figuur in de publicatie is geplot is waarschijnlijk te weinig (zie ook 3:67), de hele set aan ruwe data wellicht te veel (1:3, 1:5). Daarentegen heb je wel de primaire data nodig om te kunnen verifiëren of om nieuwe datasets op te bouwen (3:67). Geïnterviewde 16 noemt het een compromis: tijdens het onderzoek heb je genoeg opslagruimte nodig om te kunnen werken, voor archivering gebruik je een kleiner deel om bijvoorbeeld afgeleide meetdata en gebruikerstesten op te slaan (zie ook ONDERZOEKSDATA LEVELS).

Het TU beleid stelt dat alle onderzoeksdata, code en documentatie die nodig zijn om de resultaten van het onderzoek te reproduceren gearchiveerd moeten worden in het 4TU datacenter (2:78). Geïnterviewde 3 voegt daar nadrukkelijk de mogelijkheid van reuse bij (3:16, 3:17, 3:67).

Het werk wat door [master] studenten gedaan is, is vaak veel en kan niet op de gebruikelijke netwerklocaties blijven staan (vanwege het formaat). Dit wordt bij één van de geïnterviewden nu opgeslagen op harde schijven die weer worden opgeslagen in een kast op het kantoor van de supervisor[30] (11:43).

Vooral de scripts moeten gearchiveerd worden (scripts om data te genereren en om data te bewerken) (15:27)

[30] Voor deze these gaat het om data die ook daadwerkelijk in een onderzoek wordt gebruikt en daarom wordt bewaard.

Concluderend voor wat er gearchiveerd moet worden:
Alle onderzoeksdata en bijbehorende informatie die nodig zijn om de resultaten van het onderzoek te reproduceren. De onderzoeksdata en de bijbehorende beschrijving (metadata en/of datadocumentatie) op een detailniveau die de data geschikt maken voor hergebruik.

11.11.2. Archiveerbaarheid hoe

Soms wordt er gebruik gemaakt van een standaard voor metadata, dat geldt voor mensen die bij 4TU archiveren, ze gebruiken daar Dublin Core (2:46).
Bij 4TU wordt een abstract opgeslagen met een verwijzing naar de data die staat opgeslagen in een repository die in het vakgebied geaccepteerd is (3:14).
Om blind te kunnen (peer) reviewen is het in sommige vakgebieden noodzakelijk dat de data niet terug te herleiden is tot de onderzoeker (5:25,5:26).
Voor de archivering van data wordt een PID gebruikt (5:23, 7:45, 9:37, 15:17).
Er wordt ook gearchiveerd bij Zenodo (7:45). => Dit komt terug bij (Archiveerbaarheid waar).
Er wordt gearchiveerd bij het journal waarin gepubliceerd wordt, volgens de eisen die het journal stelt. [Journals zijn over het algemeen vakgebied gebonden]. => Dit komt terug bij ARCHIVEERBAARHEID WAAR.

Concluderend:
Door gebruik te maken van een PID (DOI).
Door in 4TU alleen een abstract op te slaan en te verwijzen naar een andere repository (en zo aan het beleid van de TU en van wat algemeen geaccepteerd is in het vakgebied te voldoen).
In de peer review fase kan het nodig zijn dat de data anoniem zijn gearchiveerd.

11.11.3. Archiveerbaarheid waar

Op officiële locaties:

- 4TU (2:78, 5:23, 6:21, 9:5)
- Daar waar de regelgeving van de partners vereist dat het wordt opgeslagen (3:12).
- In door het vakgebied geaccepteerde repositories (3:13).
- Zenodo (7:4)
- Bij het journal waarin gepubliceerd wordt (9:5)

Op de locatie die tijdens het onderzoek werd gebruikt:

- Het blijft staan op het rekencluster (wat dus geen archiveringsactie is) en indien nodig wordt het verhuisd naar SurfSara of LSE in München (11:27).
- Op een lokaal systeem (wat dus geen archiveringsactie is) (11:27).
- In Bit Bucket (code) (14:19).
- Op losse harde schijven die in een kast worden opgeslagen (11:43).

Bovenstaande suggereert dat de definitie van archiveren niet uitsluitend het opslaan bij een repository die nadrukkelijk bedoeld is om te archiveren betekent. Een deel van de geïnterviewden lijkt archiveren te beschouwen als het opslaan op een plek waar ze later de data terug kunnen vinden.
Dit kan betekenen dat men de data op de locatie laat staan die men tijdens het onderzoek gebruikte (zoals het rekencluster).

Concluderend:
Archiveerbaarheid wordt verschillend benaderd:

1. Het opslaan van relevante data uit het afgeronde onderzoek bij een daarvoor gespecialiseerde repository.
2. Het opslaan van voldoende relevante data uit het lopende of afgeronde onderzoek op een plek waar deze data later teruggevonden kan worden, om (delen van) het onderzoek te kunnen herhalen.

Voor dit onderzoek zijn met name de bevindingen die aansluiten op het tweede deel van belang, aangezien die nog bij de functionaliteit van het werksysteem passen (hier is nog sprake van lopend onderzoek).

11.11.4. Archiveerbaarheid eisen

Eisen over de duur van de archivering, over hoe er gearchiveerd moet kunnen worden, waar en wat er gearchiveerd wordt, staan hiervoor benoemd. Daar komt bij:

- Het archief moet voldoende capaciteit hebben om de data op te kunnen slaan (3:67, 11:43);
- Een archief moet eenvoudig in het gebruik zijn (7:5).

Ten aanzien van de eisen aan de duur kan men zeggen (geparafraseerd):

- Zo lang als het nuttig is, of dat 1, 3, 5, 10 of 100 jaar is, weet je niet van te voren (3:16).
- Een jaar of tien is een mooie termijn (11:44).
- Zolang als het artikel gepubliceerd wordt door het journal (als de data daar ook gearchiveerd staan) (9:36).
- De duur van de archivering is afhankelijk van hoe lang de publicatie gepubliceerd staat (9:36).
- Een academische carrière duurt door de bank genomen 45 jaar, vanuit dat oogpunt lijkt 50 jaar archivering redelijk (9:36).

Deze duur slaat met name op de data die gearchiveerd worden volgens het eerste deel van de definitie in de vorige paragraaf. Voor het tweede deel van de definitie wordt teruggegrepen naar de paragraaf ONDERZOEKSDATA OPRUIMEN: data blijft staan zolang er ruimte is.

Concluderend:
Het archief moet:

- voldoende capaciteit hebben om de data op te kunnen slaan;
- Een archief moet eenvoudig in het gebruik zijn.

Gearchiveerde data hebben nut zolang:

1. De data hergebruikt worden voor ander onderzoek;
2. De data worden gebruikt voor het herhalen van onderzoek;
3. De bijbehorende publicatie nog gepubliceerd staat (dit kan korter zijn dan nummer 1).

Een termijn aan archivering hangen lijkt derhalve arbitrair. De termijn zou moeten afhangen van de periode dat de data nog nut hebben.

11.12. Archiveerbaarheid wel of niet

Op geïnterviewden 8, 10 en 12 na, gaat iedereen in op archivering en past ook een vorm van archivering toe. Bij 8, 10 en 12 is er niet expliciet naar gevraagd.

11.12.1. Co-occurrence binnen de archiveerbaarheid categorie

De verwantschappen in de code categorie archivering zijn het sterkst van allemaal. Ze staan samengevat in TABEL 87.

Verwantschappen binnen de volgbaarheid categorie			
	Code 1	Code 2	Coëfficiënt
1	Archief ja	Archief eisen	0.50, 6
2	Archief ja	Archief waar	0.46, 7
3	Archief eisen	Archief hoe	0.40, 5
4	Archief eisen	Archief waar	0.31, 6
5	Archief ja	Archief hoe	0.30, 5
6	Archief wat	Archief waar	0.26, 2
7	Archief eisen	Archief wat	0.25, 4

Uitleg:

- Er is grote verwantschap tussen de verschillende archiveringscodes. Met name de archiveringseisen hebben verwantschap met alle andere codes. Dit suggereert dat de eisen wellicht bepalend zijn voor hoe, wat en waar er gearchiveerd wordt. Dat klopt ook met de formulering van de eisen (ARCHIVEERBAARHEID EISEN).
- Als men daadwerkelijk archiveert, heeft dat vooral verwantschap met de eisen waaraan het archief moet voldoen, waar men archiveert en hoe men dat doet. Het heeft weinig verwantschap met wat wordt gearchiveerd. Een verklaring zou kunnen zijn dat verschillende onderzoekers verschillend denken over wat er moet worden gearchiveerd.
- Wat wordt gearchiveerd lijkt wel verwantschap te hebben met waar wordt gearchiveerd. Dit komt met name door de verplichting in het beleid om in het 4TU datacenter op te slaan en de gewoonte van één van de onderzoekers om data vooral zoveel en zolang mogelijk op het rekencluster te laten staan. Eigenlijk zijn er maar twee bronnen, wat deze verwantschap onvoldoende sterk maakt om te laten bestaan.

Tabel 87, Verwantschappen binnen de volgbaarheid categorie

11.12.2. Archiveerbaarheid conclusie

De definitie van archiveren wordt verbreed. Archiveren kan opgesplitst worden in twee vormen:

3. Het opslaan van relevante data uit het afgeronde onderzoek bij een daarvoor gespecialiseerde repository.
4. Het opslaan van voldoende relevante data uit het lopende of afgeronde onderzoek op een plek waar deze data later teruggevonden kan worden, om (delen van) het onderzoek te kunnen herhalen.

Voor dit onderzoek zijn met name de bevindingen die aansluiten op het tweede deel van belang, aangezien die nog bij de functionaliteit van het werksysteem passen (hier is nog sprake van lopend onderzoek).

Wat van de bevindingen i.i.g. voor het tweede deel van de definitie geldt:
Wat moet worden gearchiveerd:

- Alle onderzoeksdata en bijbehorende informatie die nodig zijn om de resultaten van het onderzoek te reproduceren;
- De onderzoeksdata en de bijbehorende beschrijving (metadata en/of datadocumentatie) op een detailniveau dat het geschikt maakt voor hergebruik.

Waar wordt gearchiveerd:

- Data worden wel bewaard, maar niet actief gearchiveerd (veelal blijft het staan op de locatie die ook tijdens het onderzoek gebruikt werd).

Het archief moet:

- voldoende capaciteit hebben om de data op te kunnen slaan;
- Een archief moet eenvoudig in het gebruik zijn.

Gearchiveerde data hebben nut zolang:

- De data hergebruikt worden voor ander onderzoek;
- De data worden gebruikt voor het herhalen van (delen van het) onderzoek.

Data kan blijven staan zolang er voldoende ruimte is op de gebruikte opslaglocatie.

11.13. Ontwerp

Onder ontwerp zijn alle ideeën verzameld over het inrichten van RDM. In de interviews zijn ideeën geopperd over het ontwerp in het algemeen, over share and sync invulling, en over ELN's (TABEL 88). Dit wordt hieronder één voor één uitgewerkt.

Subcodes in de codecategorie ontwerp	
Subcode omschrijving	Subcode in Atlas
Ontwerp in het algemeen	archit ontwerp
Ontwerp van een sync and share oplossing	archit share and sync
Ontwerp van een ELN oplossing	archit eln

Tabel 88, Subcodes in de codecategorie ontwerp

11.13.1. Ontwerp algemeen

Geïnterviewde 2 geeft aan dat het goed is om wat algemene oplossingen te hebben, maar dat je voor de specifieke gevallen een specifieke oplossing moet zien te vinden. Het is onmogelijk om aan alle behoeften te voldoen (2:84).

1. Geïnterviewde 7 voegt daar aan toe. "Een monolithisch programma dat alles doet gaat nooit lukken" (7:20).
2. Geïnterviewde 7 zegt te hechten aan de Unix filosofie: "Maak je software uit kleine deeltjes die allemaal één ding heel goed doen. Maar die ook heel makkelijk met de andere stukken kunnen communiceren zodat jij dan ook flexibiliteit hebt." Dat zou je kunnen herformuleren naar: kleine diensten die gemakkelijk gecombineerd kunnen worden tot één grotere dienst (geïnterviewde 7 noemt voor zo'n grotere dienst het voorbeeld van een ELN, 7:21).
3. Geïnterviewde 2 zegt ten slotte dat als iets eenvoudig te gebruiken is en op een eenvoudige manier wordt aangeboden (er wordt als voorbeeld een template genoemd) dan is de kans dat iets gebruikt wordt een stuk groter (2:24).

Vertaald naar RDM oplossingen kunnen de uitspraken als volgt worden geformuleerd:

Een (onderdeel van) een RDM oplossing moet generiek genoeg zijn om het deelgebied af te dekken, maar moet ook eenvoudig aangevuld kunnen worden met specifiekere functionaliteit door:

- Kleine specifieke diensten te gebruiken;
- Door templates te gebruiken.

Geïnterviewde 11 heeft twee problemen waar hij graag dezelfde oplossing voor wil. Master studenten leveren hun data meestal in op een harde schijf. Hij heeft er inmiddels een kast vol mee. In onderzoeken die hij begeleid wordt vaak met heel veel data gewerkt die ofwel van een harde schijf ofwel van tape moeten komen. In beide gevallen is de wachttijd groot. Hij zou graag opslag willen hebben met voldoende capaciteit waar hij eenvoudig al deze data op zou kunnen zetten en vanaf zou kunnen halen. Een archief dus, maar niet voor data die bij een publicatie horen (11:43, 11:48, 11:50). Veel van die data staat, zolang er ruimte is, op het HPC cluster (11:16). Dat is een gebruik dat meerdere geïnterviewden hanteren (1:38, 1:40, 8:6, 9:30, 14:10, 14:29, 15:28).
Dit kan samengevat worden naar:

Er is behoefte aan eenvoudig toegankelijke online opslag met een snelle responstijd[31] waar grote hoeveelheden data van oude onderzoeken, maar ook van Master theses opgeslagen kunnen worden.

[31] In ieder geval sneller dan die van tape.

11.13.2. Ontwerp share and sync

De conclusie van deze code wordt het best samengevat door geïnterviewden 2 en 3.

- " (Project drive should have the Dropbox functionality in the end?) Speaker 2: Of course that's what they need." (2:36).
- "Dus mocht er nu een manier zijn dan dat je gewoon kan markeren, share deze folder even, bij voorkeur vanuit de Unix environment, share deze folder even met deze persoon dan zal ik dat gebruiken." (3:53)
- "Van is er een eenvoudige manier zodat ik gewoon deze folder kan markeren om te delen stel je voor de hele TU storage is een gigantische Dropbox waarbij je kan zeggen dat je deze folder deelt die met die en die mensen." (3:54)

Er is behoefte aan grote opslag met share and sync functionaliteit.

Groot is niet direct gekwantificeerd, maar geïnterviewde 3 zegt dat 250Gb te weinig is (3:52) en geïnterviewde 11 lijkt het acceptabel te vinden dat 8Tb als basis wordt beschouwd en alles wat meer is verantwoord moet worden in een aanvraag (11:50).

Geïnterviewde 1 voegt daar aan toe dat hij eerder Dropbox heeft gebruikt om data tussen HPC en zijn laptop te synchroniseren (1:41). Dit zou generieker geformuleerd kunnen worden naar: de mogelijkheid om data vanaf externe datahouders te synchroniseren naar centrale opslag met behulp van share and sync software.

Op basis hiervan zou dit opnieuw geformuleerd kunnen worden naar[32]:

- Er is behoefte aan opslag met share and sync functionaliteit, met een capaciteit van minimaal 8Tb per onderzoek.
- Data vanaf externe datahouders moet met deze opslag gesynchroniseerd kunnen worden.

11.13.3. Ontwerp ELN

Bij de opmerkingen over ELN's worden dezelfde specificaties aangegeven als bij de algemene ontwerp. Een ELN moet opgebouwd zijn uit een aantal basiselementen aan te vullen met specifieke functionaliteit uit softwarediensten en/of templates (2:27, 7:22, 7:46, 7:48, 7:57, 7:58, 12:36).

Een specifieke invulling voor RDM wordt samengevat door geïnterviewde 1:

"So you would imagine kind of like an electronic journal that says: Today I ran this calculation and the input file was this, the output file was that? That could be something. Yeah, yeah, yeah, yeah. And that you store your data somewhere and you can link to it. You give the data of course some unique name." (1:55).

Het kunnen documenteren wat je hebt gedaan, maar ook waarom je bepaalde keuzes hebt gemaakt komt terug bij verschillende geïnterviewden (1:55, 7:16, 12:36).

Dat geldt ook voor het kunnen linken naar datasets. Hiermee wordt bedoeld dat het mogelijk is een verwijzing in een ELN op te nemen naar een dataset. Hierbij wordt de suggestie gewekt dat dit clickbare verwijzingen zouden moeten zijn (1:57, 7:58). Geïnterviewde 12 gebruikt een template met

[32] Deze oplossing kan ook gebruikt worden voor de specifieke vraag van geïnterviewde 11 in de ontwerp algemeen paragraaf.

databeschrijvingen dat door de meetopstelling gevuld kan worden en in het labnotebook wordt opgenomen (12:36).

Concluderend kan gesteld worden:

Er is behoefte aan een ELN met basisfunctionaliteiten als de mogelijkheden om:

- Te kunnen documenteren wat je hebt gedaan;
- Te kunnen documenteren waarom je bepaalde keuzes hebt gemaakt;
- Te kunnen linken naar datasets;
- Het kunnen toevoegen van bepaalde templates (bijvoorbeeld van databeschrijvingen).

Bovendien zou het ELN uitgebreid moeten kunnen worden met specifieke onderzoeksgerelateerde functionaliteit.

11.13.4. Ontwerp eisen naar aanleiding van het gedrag van onderzoekers.

Het is waarschijnlijk verstandig om hier ook de eerder genoemde onderzoeksgewoonten te vertalen naar specificaties voor ontwerp (zie Werkwijze).

- RDM tooling dient ondersteunende werkzaamheden uit handen te nemen van de wetenschapper zodat hij/zij zich kan bezighouden met het onderzoek.
- RDM tooling moet per situatie zo aangepast kunnen worden dat het het best aansluit bij de werkzaamheden.
- RDM tooling moet zo min mogelijk tijd vergen om aan te leren.

11.13.5. Co-occurrence binnen de ontwerp code categorie

Er is geen relevante verwantschap tussen de subcodes.

11.13.6. Ontwerp conclusie

Een (onderdeel van een) RDM oplossing moet generiek genoeg zijn om het deelgebied af te dekken, maar moet ook aangevuld kunnen worden met specifiekere functionaliteit door:

- Kleine specifieke diensten te gebruiken;
- Door templates te gebruiken;
- RDM tooling dient ondersteunende werkzaamheden uit handen te nemen van de wetenschapper zodat hij/zij zich kan bezighouden met het onderzoek;
- RDM tooling moet per situatie zo aangepast kunnen worden dat het het best aansluit bij de werkzaamheden;
- RDM tooling moet zo min mogelijk tijd vergen om aan te leren.

Specifiek voor een share and sync oplossing:

- Er is behoefte aan opslag met share and sync functionaliteit, met een capaciteit van minimaal 8Tb per onderzoek.
- Data vanaf externe datahouders moet met deze opslag gesynchroniseerd kunnen worden.

Specifiek voor een ELN, Er is behoefte aan een ELN met basisfunctionaliteiten als de mogelijkheden om:

- Te kunnen documenteren wat je hebt gedaan;
- Te kunnen documenteren waarom je bepaalde keuzes hebt gemaakt;

- Te kunnen linken naar datasets;
- Het kunnen toevoegen van bepaalde templates (bijvoorbeeld van databeschrijvingen).

Bovendien zou het ELN uitgebreid moeten kunnen worden met specifieke onderzoeksgerelateerde functionaliteit.

Specifiek voor grote hoeveelheden data:

Er is behoefte aan eenvoudig toegankelijke online opslag met een snelle responstijd[33] waar grote hoeveelheden data van oude onderzoeken, maar ook van Master theses opgeslagen kunnen worden.

[33] In ieder geval sneller dan die van tape.

12. Bijlage co-occurrence tussen de categorieën

De sterkte van een verwantschap (co-occurrence) tussen twee codes wordt uitgelegd in BIJLAGE UITLEG CO-OCCURRENCE/VERWANTSCHAP. In TABEL 89 staan de verwantschappen tussen de verschillende code-paren tussen verschillende categorieën samengevat. Wat verwantschap is en welke overwegingen hierbij zijn gemaakt staat in dezelfde bijlage. De verwantschappen worden in dezelfde volgorde behandeld als in de BIJLAGE UITWERKING INTERVIEWS.

Co-occurrence tussen de categorieën	bvlg contract	dd data bewerken	dd data verzamelen	dd inhoud	dd link	del code	md fdst groep	md fdst persoonlijk	md nmc persoonlijk	ont eln	opsl code	rd eln	rd gevoelig	vind folderstructuur	vind naamconventie	vind notebook	vlg naamconventie
bvlg contract													0.25				
dd data bewerken												0.40					
dd data verzamelen												0.29					
dd inhoud												0.25					
dd link										0.27						0.33	
del code											0.35						
md fdst groep														0.27			
md fdst persoonlijk														0.21			
md nmc persoonlijk																	0.22
ont eln					0.27												
opsl code						0.35											
rd eln		0.40	0.29	0.25													
rd gevoelig	0.25																
vind folderstructuur							0.27	0.21									
vind naamconventie																	0.25
vind notebook					0.33												
vlg naamconventie									0.22						0.25		

Opmerkingen:
Alle waarden staan twee keer in de tabel (de tabel is symmetrisch t.o.v. de diagonaal).
Legenda:
Waarden van 0.2 tot 0.3
Waarden van 0.3 tot 0.4
Waarden van 0.4 tot 0.5

Tabel 89, Co-occurence tussen de categorieën

12.1. Cross-categorie verwantschap Onderzoeksdata

In de categorie onderzoeksdata bestaan vier verwantschappen met codes uit andere categorieën die relevant zijn in sterkte. Ze staan samengevat in onderstaande TABEL 90.

Verwantschappen voor onderzoeksdata			
	Subcode onderzoeksdata	Subcode andere categorie	
1	Onderzoeksdata ELN	Datadocumentatie data bewerken	0.4
2	Onderzoeksdata ELN	Data documentatie data verzamelen	0.29
3	Onderzoeksdata ELN	Data documentatie inhoud	0.25
4	Onderzoeksdata gevoelig	Data beveiliging Onderzoeksdata ELN	0.25

1. In de datadocumentatie, in de vorm van een lablogboek, bijhouden welke bewerkingen de data ondervinden, het één van papier en het ander elektronisch.
2. Twee voorbeelden waarbij verzamelde data in een lablogboek terechtkomen, het één van papier en het ander elektronisch.
3. Hoe datadocumentatie in een lablogboek de onderzoeksdata inhoudelijk beschrijven.
4. Specifiek voor het geval dat voor bepaalde onderzoeksdata een 'geheimhoudings'contract getekend moet worden.

Tabel 90, Verwantschappen voor onderzoeksdata

Concluderend kan gesteld worden dat er specifieke voorbeelden bestaan waarbij wetenschappers op de TU Onderzoeksdata in een lablogboek beschrijven hoe ze data verzamelen, bewerken en hoe de data er inhoudelijk uitzien.

T.a.v. de omgang met gevoelige onderzoeksdata is het soms nodig een geheimhoudingscontract te tekenen.

12.2. Cross-categorie verwantschap metadata

In de categorie metadata bestaan drie verwantschappen met codes uit andere categorieën die relevant zijn in sterkte (TABEL 91).

Verwantschappen voor metadata			
	Subcode onderzoeksdata	Subcode andere categorie	
1	Metadata folderstructuur groep	Vindbaarheid folderstructuur	0.27
2	Metadata folderstructuur persoonlijk	Vindbaarheid folderstructuur	0.21
3	Metadata naamconventie persoonlijk	Volgbaarheid naamconventie	0.22

1. Bij bepaalde onderzoeksgroepen wordt een vaste folderstructuur gebruikt waar men afspreekt wat in welke folder moet worden opgeslagen. Dit bevordert de vindbaarheid van de data.
2. Hetzelfde geldt voor mensen die persoonlijk een folderstructuur opzetten.
3. Door het toevoegen van versienummers, een datum, bepaalde parameters van het experiment (bijvoorbeeld een snelheid) in de naam van het bestand, worden de opeenvolgende databewerkingen duidelijker.

Tabel 91, Verwantschappen voor metadata

Concluderend kan gesteld worden dat folderstructuren vooral helpen bij de vindbaarheid van de onderzoeksdata en dat naamconventies helpen bij het volgen van de databewerkingen.

12.3. Cross-categorie verwantschap datadocumentatie

In de categorie data documentatie bestaan vijf verwantschappen met codes uit andere categorieën die relevant zijn in sterkte. Drie ervan zijn besproken in **ERROR! REFERENCE SOURCE NOT FOUND.** maar worden voor de volledigheid hier weer genoemd (TABEL 92).

Verwantschappen voor datadocumentatie			
	Subcode onderzoeksdata	Subcode andere categorie	
1	Data documentatie link	Vind Notebook	0.33
2	Data documentatie link	Ontwerp ELN	0.27
1. Doordat de datadocumentatie verwijst naar de datalocatie, is het mogelijk de data eenvoudig terug te vinden. 2. Datadocumentatie zou moeten kunnen linken naar de beschreven data, dit wordt genoemd in combinatie met het gebruik van lablogboeken.			

Tabel 92, Verwantschappen voor datadocumentatie

De lablogboeken worden gebruikt om te linken naar de datalocatie, of het is een wens om dit toe te kunnen passen, waardoor de onderzoeksdata eenvoudig is terug te vinden.

12.4. Cross-categorie verwantschap data opslag en delen code

In de categorieën data opslag en delen data bestaat één verwantschap die relevant is in sterkte (TABEL 93).

Verwantschap data opslag en data delen			
	Subcode onderzoeksdata	Subcode andere categorie	
1	Data opslag code	Data delen code	0.35
1. Specifiek bedoelde systemen voor de opslag van code hebben ook de mogelijkheid om data te delen met derden. Ze worden daar dan ook voor gebruikt.			

Tabel 93, Verwantschap data opslag en data delen

12.5. Cross-categorie verwantschap vindbaarheid en volgbaarheid

In de categorieën vindbaarheid en volgbaarheid bestaat één verwantschap die relevant is in sterkte (TABEL 94).

Verwantschap vindbaarheid en volgbaarheid			
	Subcode onderzoeksdata	Subcode andere categorie	
1	Vindbaarheid naamconventie	Volgbaarheid naamconventie	0.25
1. Specifiek genoemd waarbij de naam wordt voorzien van experimentele omstandigheden zoals snelheid of temperatuur en de naam bovendien wordt voorzien van een datum of versienummer. De omstandigheden en de datum kunnen helpen in de vindbaarheid en het versienummer of de datum kunnen ook helpen bij de volgbaarheid.			

Tabel 94, Verwantschap vindbaarheid en volgbaarheid

Concluderend kan gesteld worden dat door het toevoegen van versienummers, een datum, bepaalde parameters van het experiment (bijvoorbeeld een snelheid) in de naam van het bestand, het volgen van de opeenvolgende databewerkingen makkelijker. Het kan het zoeken op onderzoeksdata bovendien wat eenvoudiger maken.

12.6. Conclusie verwantschappen cross-categorie

Er zijn drie belangrijke invloeden op vindbaarheid en volgbaarheid, een ELN, folderstructuren en naamconventies:

In een ELN kan men een verwijzing maken (bijvoorbeeld een snelkoppeling) naar een dataset. Dat maakt terugvinden eenvoudiger. Door per onderzoek een bekende folderstructuur te gebruiken en die consequent te hanteren, kan men relevante data eenvoudig terugvinden.

Als men in het ELN de onderzoeksdata beschrijft, en ook beschrijft hoe deze data zijn verzameld en welke bewerkingen ze hebben ondergaan, dan sluit dit aan op de definitie van volgbaarheid. Door de databestanden die bij een onderzoeksstap horen te voorzien van bepaalde beschrijvende parameters (denk aan een snelheid in het experiment, of de datum dat de onderzoeksstap werd uitgevoerd) dan kan daar op gezocht worden (vindbaarheid) maar het kan ook helpen met volgbaarheid (als er elke datum een andere snelheid wordt gebruikt bijvoorbeeld).

13. Bijlage Uitleg Co-Occurrence/verwantschap

Het is interessant om te zien of bepaalde codes elkaar beïnvloeden. Als bijvoorbeeld blijkt dat de relatie tussen het gebruiken van een naamconventie en het terug kunnen vinden van data sterk is, dan zou je daar uit voorzichtig kunnen concluderen dat naamconventies helpen bij het terugvinden van data.

De gebruikte tooling (Atlas.ti versie 8) heeft hier een hulpmiddel voor, de code co-occurrence table. In een Code Co-Occurrence (cooc) table, worden de verschillende gebruikte codes onderzocht op hun samenhang. Dit gaat als volgt:

- In elk interview worden belangrijke uitspraken als quote gelabeled;
- Elke quote krijgt 1 of meer codes;
- De codes worden onderzocht op gemeenschappelijk onderliggende quotes;
- Hier wordt een genormaliseerd getal voor berekend (de c-coefficient) in Atlas.ti, dat een soort correlatie aangeeft.

Hoewel de coëfficiënt hetzelfde gedrag vertoont als die van Spearman of Pearson voor correlatie, is het niet hetzelfde en is het ook niet mogelijk er een mate van significantie aan af te meten. De ene onderzoeker zal immers heel anders coderen dan de ander (bijvoorbeeld wat minder grotere quotes vs. veel kleine quotes). Bovendien gaat het in dit onderzoek om een populatie van 14 personen, die een semigestructureerd interview werd afgenomen. De coëfficiënt laat wel het onderlinge verschil zien tussen de code paren (Anon 2015).

Er zijn in totaal 71 codes die met elkaar vergeleken kunnen worden. Dat betekent dat er 71 boven 2 codeparen zijn (2485). Deze code co-occurrences kunnen geëxporteerd worden uit Atlas. TI. Eeen deel van deze tabel voor de code categorieën onderzoeksdata (rd), vindbaarheid (vind) en volgbaarheid (volg) staat in TABEL 95. Alle databewerkingen op de gehele tabel en het resultaat worden in de komende pagina's beschreven.

Een waarde van de c-coefficient die behoorlijke verwantschap tussen codes weergeeft, lijkt een stuk hoger te moeten liggen. Dit wordt in de volgende paragraaf bepaald.

13.1. Significantie van de c-coefficient

Om te kunnen bepalen welke waardes van de c-coefficient voor dit onderzoek een relevante verwantschap vertegenwoordigen, wordt er in deze paragraaf dieper ingegaan op de getallen uit de tabel. Het enige waar van uit kan worden gegaan ten aanzien van de codes en coëfficiënten die er zijn, zijn de waardes die uit Atlas geëxporteerd kunnen worden. Coderen is iets persoonlijks en bovendien iets wat inherent niet op één manier uitgevoerd kan worden, maar op oneindig veel manieren. Er wordt dus naar de eigen codes gekeken om te bepalen welke verwantschappen sterker zijn en welke niet meegeteld worden. Dat gebeurt in een aantal stappen.

Ten eerst als TABEL 95 in ogenschouw wordt genomen lijkt het dat de meeste hogere waarden, zich op het oog concentreren binnen een categorie (bijvoorbeeld vindbaarheid met de subcodes vind folderstructuren, vind ja, vind naamconventie en vind notebook). Dit kan gecontroleerd worden door de tabel op te splitsen met alleen de categoriewaarden (de 'diagonaal' **dik rood omrand in de tabel**), en het omgekeerde, dus de verwantschap van codes met codes uit een andere categorie. Als de gedachtegang klopt, zullen de relaties binnen de diagonaal relatief hogere waarden hebben dan buiten de diagonaal.

	rd actief	rd code	rd definitie	rd eln	rd gevoelig	rd herkomst	rd levels	rd meetdata	rd opruimen	rd simulatie	vind folderstructuur	vind ja	vind naamconventie	vind notebook	vlg code	vlg folderstructuur	vlg ja	vlg naamconventie	vlg transformatie	vlg versie
rd actief		0.00	0.07	0.00	0.00	0.04	0.02	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rd code	0.00		0.28	0.00	0.00	0.07	0.09	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rd definitie	0.07	0.28		0.00	0.05	0.14	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
rd eln	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.07	0.00
rd gevoelig	0.00	0.00	0.05	0.00		0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rd herkomst	0.04	0.07	0.14	0.00	0.04		0.07	0.27	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00
rd levels	0.02	0.09	0.05	0.00	0.00	0.07		0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.04	0.00
rd meetdata	0.04	0.00	0.02	0.00	0.04	0.27	0.06		0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00
rd opruimen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.03	0.00	0.00	0.00
rd simulatie	0.00	0.04	0.00	0.00	0.00	0.12	0.00	0.04	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vind folderstructuur	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.38	0.04	0.00	0.00	0.08	0.04	0.03	0.00	0.00
vind ja	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38		0.17	0.26	0.00	0.08	0.12	0.12	0.03	0.02
vind naamconventie	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.17		0.06	0.00	0.18	0.05	0.25	0.03	0.07
vind notebook	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.06		0.00	0.08	0.05	0.05	0.06	0.00
vlg code	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00		0.00	0.22	0.04	0.15	0.31
vlg folderstructuur	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08	0.18	0.08	0.00		0.10	0.21	0.07	0.00
vlg ja	0.00	0.00	0.00	0.08	0.00	0.01	0.02	0.01	0.03	0.00	0.04	0.12	0.05	0.05	0.22	0.10		0.15	0.49	0.20
vlg naamconventie	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.12	0.25	0.05	0.04	0.21	0.15		0.09	0.10
vlg transformatie	0.00	0.00	0.00	0.07	0.00	0.01	0.04	0.02	0.00	0.00	0.00	0.03	0.03	0.06	0.15	0.07	0.49	0.09		0.20
vlg versie	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.07	0.00	0.31	0.00	0.20	0.10	0.20	

Legenda
0
0 tot 0.1
0.1 tot 0.2
0. 2tot 0.3
0.3 tot 0.4
0.4 tot 0.5

Tabel 95, code co-occurrence tabel voor onderzoeksdata, vindbaarheid en volgbaarheid

Stap 1, samenvatten over de totale co-occurrence tabel

In eerste instantie worden de waardes gegeven over de gehele co-occurrence tabel (TABEL 96 en TABEL 97):

Verdeling codeparen per coëfficiënt ruwe data		
Kleur verdeling	Aantal paren	percentage
0 precies	3940	79%
> 0 tot 0.1	818	16%
0.1 tot 0.2	134	3%
0.2 tot 0.3	50	1%
0.3 tot 0.4	14	0%
0.4 tot 0.5	14	0%
Totaal aantal codeparen	4970	

Tabel 96, Verdeling codeparen[34] per coëfficiënt ruwe data

Centrummaten en maximum ruwe data	
Centrummaten en maximum	waarde
mean	0.01
median	0.00
modus	0
max	0.50

Tabel 97, Centrummaten en maximum ruwe data

Tabel 96 laat zien dat 95% van alle waarden een coëfficiënt van 0.1 of kleiner heeft.

Stap 2, splitsen van waardes binnen en tussen code categorieën

De waardes van TABEL 96 en TABEL 97 worden nog eens bepaald voor de waardes binnen de code categorieën en voor de waardes tussezn verschillende code categorieën. Als de aanname klopt, liggen de sterkere waarden er binnen.

Binnen de code categorieën:

Verdeling codeparen per coëfficiënt binnen categorie		
Kleur verdeling	Aantal paren	percentage
0 precies	176	43%
> 0 tot 0.1	126	31%
0.1 tot 0.2	48	12%
0.2 tot 0.3	34	8%
0.3 tot 0.4	10	2%
0.4 tot 0.5	12	3%
Totaal aantal overgebleven codes	406	

Tabel 98, Verdeling codeparen per coëfficiënt binnen categorie na verwijdering 0 waarden

De statistische waarden worden:

Centrummaten en maximum binnen categorie	
Centrummaten en maximum	waarde
mean	0.07
median	0.03
modus	0
max	0.50

Tabel 99 Centrummaten en maximum binnen categorie

[34] De tabel laat een dubbel aantal paren zien, omdat elke codepaar twee keer voorkomt, dus bijvoorbeeld vind folderstructuur en vind ja, maar ook andersom vind ja en vind folderstructuur.

Het gemiddelde gaat omhoog. De meest voorkomende waarde (modus) is nog steeds 0. In 43% van de gevallen is er geen verwantschap. Dat betekent dat er fors meer verwantschap is dan in de tabellen over alle waarden (TABEL 96, TABEL 97) waar dit percentage bijna het dubbele was (79%). Dit ondersteunt de aanname dat rond de diagonaal de meeste verwantschap optreedt.

Tussen de code categorieën

Verdeling codeparen per coëfficiënt buiten categorie		
Kleur verdeling	Aantal paren	percentage
0 precies	3764	82%
> 0 tot 0.1	692	15%
0.1 tot 0.2	86	2%
0.2 tot 0.3	16	0%
0.3 tot 0.4	4	0%
0.4 tot 0.5	2	0%
Totaal aantal overgebleven codes	4564	

Tabel 100, Verdeling codeparen per coëfficiënt buiten categorie na verwijdering 0 waarden

Centrummaten en maximum buiten categorie	
Centrummaten en maximum	waarde
mean	0.01
median	0.00
modus	0
max	0.40

Tabel 101, Centrummaten en maximum buiten categorie

Het gemiddelde in TABEL 101 t.o.v. TABEL 97 verandert niet. Er zijn in TABEL 100 relatief iets meer 0 waarden dan in TABEL 96.
Dit ondersteunt de aanname dat de sterkste verwantschap in de sheet in dezelfde code categorie zit (rond de diagonaal).

Stap 3 het verwijderen van de 0 waarden

Uitgaande van de aanname dat binnen de categorieën de sterkste verwantschappen bestaan, kan nu uitgezocht worden, welke waarde een voldoende sterkte vertegenwoordigt. Eerst worden de nulwaarden in de co-occurrence tabel van de code categorieën verwijderd, zij vertegenwoordigen geen verwantschap.

Verdeling codeparen per coëfficiënt binnen categorie na verwijdering 0 waarden		
Kleur verdeling	Aantal paren	percentage
0 precies	0	0%
> 0 tot 0.1	126	55%
0.1 tot 0.2	48	21%
0.2 tot 0.3	34	15%
0.3 tot 0.4	10	4%
0.4 tot 0.5	12	5%
Totaal aantal overgebleven codes	230	

Tabel 102, Verdeling codeparen per coëfficiënt binnen categorie na verwijdering 0 waarden

Centrummaten en maximum binnen categorie na verwijdering 0 waarden	
Centrummaten en maximum	waarde
mean	0.13
median	0.08
modus	0.04
max	0.50

Tabel 103, Centrummaten en maximum buiten categorie na verwijdering 0 waarden

76% van de waarden zit onder de 0.2 (TABEL 102), de waarden schuiven dus wat op, dat blijkt ook uit het gemiddelde, de mediaan en de modus, zie TABEL 103.

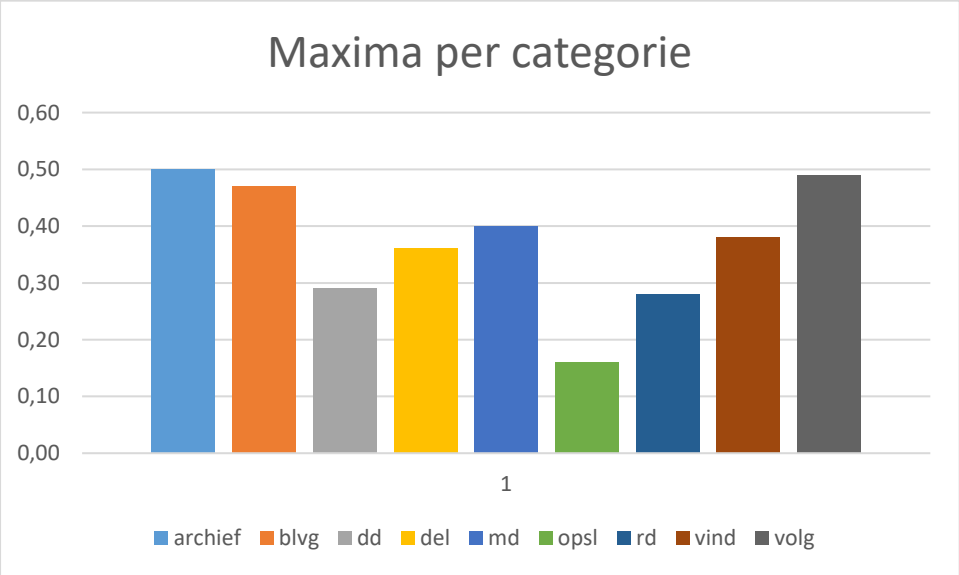
Stap 4 het bepalen van de sterkte van coëfficiënten

Als nu de statistische waarden per categorie worden vastgesteld, kan wellicht een beter beeld verkregen worden van de waarden die voldoende sterk zijn om een verwantschap te tonen. De sterkste waarden per categorie zijn de maxima. Die worden geplot voor archiveerbaarheid, security, datadocumentatie, data delen, metadata, opslag, research data, vindbaarheid en volgbaarheid. De categorieën ontwerp en individuele motivatie worden buiten beschouwing gelaten omdat die gedurende de interviewserie pas zijn ontstaan en/of niet bij iedereen zijn uitgevraagd. Ook integriteit en logging worden buiten beschouwing gelaten omdat het op zichzelf staande codes zijn en volgens Atlas.ti altijd de verwantschap 0 hebben. Dit leidt tot TABEL 104 en FIGUUR 21.

Centrummaten en maxima per categorie na verwijdering van de 0 waarden										
	ar-chief	blvg	dd	del	md	opsl	rd	vind	volg	**Gemiddeld**
mean	0.27	0.11	0.15	0.11	0.14	0.06	0.08	0.15	0.16	0.14
median	0.28	0.05	0.15	0.08	0.10	0.05	0.05	0.08	0.15	0.11
modus	0.4	0.02	0.15	0.04	0.03	0.03	0.04	0.03	0.04	0.09
max	0.50	0.47	0.29	0.36	0.40	0.16	0.28	0.38	0.49	0.37

Tabel 104, Centrummaten en maxima per categorie na verwijdering van de 0 waarden

In FIGUUR 21 kan men zien dat één waarde onder de 0.2 ligt en dat er twee vlak onder de 0.3 liggen (0.28 en 0.29). De overige zes waarden zijn hoger dan 0.3.



Figuur 21, maxima per categorie

De sterkste coëfficiënten binnen de code categorieën, daar waar de meeste verwantschap te verwachten is, liggen tussen de 0.3 en 0.5. Het lijkt dus aannemelijk dat als we elders coëfficiënten

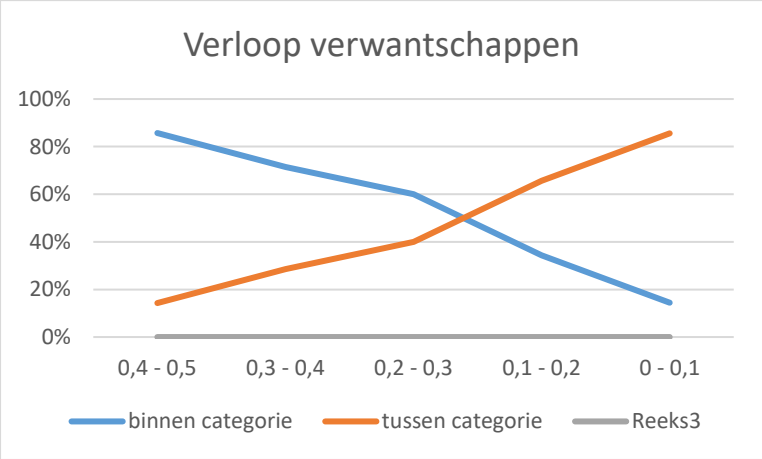
met een waarde van 0.3 en hoger vinden, dat dit een daadwerkelijk bestaande verwantschap aanduidt. Coëfficiënten tussen de 0.2 en 0.3 zijn misschien ook sterk genoeg (gezien de waarden van 0.28 en 0.29). Op basis hiervan is het eerste uitgangspunt dat coëfficiënten van 0.3 en hoger waarschijnlijk een verwantschap uitdrukken en coëfficiënten tussen de 0.2 en 0.3 mogelijk een verwantschap uitdrukken. Voor de coëfficiënten binnen de categorieën geldt dat er 28 groter dan 0.2 zijn (waarvan 11 groter dan 0.3) zie TABEL 102.

Controle van het voorgaande

Een andere manier om naar de waarden te kijken is (TABEL 98 en TABEL 100):

- Voor de coëfficiënten tussen de 0.4 tot 0.5 geldt dat 1 van de 7 waarden niet binnen een categorie ligt (ruim minder dan de helft).
- Voor de coëfficiënten tussen de 0.3 tot 0.4 geldt dat 2 van de 7 waarden niet binnen een categorie liggen (ruim minder dan de helft).
- Voor de coëfficiënten tussen de 0.2 tot 0.3 geldt dat 10 van de 25 waarden niet binnen een categorie liggen (minder dan de helft).
- Voor de coëfficiënten tussen de 0.1 tot 0.2 geldt dat 44 van de 67 waarden liggen niet binnen een categorie liggen (meer dan de helft).
- Voor de coëfficiënten tussen de van iets meer dan 0 tot 0.1 geldt dat 350 van de 409 waarden niet binnen een categorie liggen (ruim meer dan de helft).

Met de aanname van sterkere verwantschap binnen een categorie kan men concluderen:
De coëfficiënten van 0.3 en hoger worden het sterkst vertegenwoordigd binnen de categorieën.
De coëfficiënten van 0.2 en lager worden het sterkst vertegenwoordigd buiten de categorieën.
Tussen 0.2 en 0.3 is het bijna gelijk
De sterke coëfficiënten lijken 0.3 en hoger te zijn, de zwakke 0.2 en lager en 0.2 tot 0.3 is een overgangsgebied.
Dit bevestigt het eerdere uitgangspunt.
Dit staat samengevat in FIGUUR 22.



Figuur 22, verloop verwantschappen

13.1.1. Aanvullende opmerkingen sterkte verwantschappen

Bij de kwalitatieve uitwerking van de verwantschappen wordt er vanuit gegaan dat boven de 0.3 inderdaad bestaande verwantschappen zijn en dat tussen 0.2 en 0.3 mogelijk verwantschap bestaat. Waarden onder de 0.2 worden genegeerd. Voor gevallen tussen de 0.2 en 0.3 zou men eventueel kunnen kijken hoeveel geïnterviewden deze verwantschap hebben geraakt in hun interviews. Als dat

er weinig zijn, is het mogelijk een zwakkere code. Om die reden worden alle codeparen binnen de categorieën hieronder samengevat in TABEL 105. Die erbuiten worden samengevat in TABEL 106.

Als de categorie met de sterkste verwantschappen (archief) als voorbeeld genomen wordt, dan komen deze verwantschappen voor in twee (archief wat met archief waar) tot zeven (archief ja archief waar) verschillende interviews. Het aantal interviews waarin verwantschappen worden genoemd per codepaar van archief is: 2, 4, 5, 5, 6, 6, 7. Twee lijkt een erg laag aantal te zijn, tenzij het om een heel specifiek onderwerp gaat. Voor waarden tussen de 0.2 en 0.3 die in vier of meer interviews besproken worden, wordt uitgegaan van redelijke kans op verwantschap. Voor twee of drie interviews wordt uitgegaan van een lage kans op verwantschap tenzij het om een heel specifiek onderwerp gaat, dat in slechts enkele interviews besproken is. Zo gebruikt bijvoorbeeld lang niet iedereen een vorm van ELN en gebruikt ook lang niet iedereen een rekencluster. Om dit te onderzoeken zijn TABEL 105 en TABEL 106 gemaakt.

Uit TABEL 105:

- dd eln en dd link: specifieke verwantschap waarin wordt beschreven hoe je een ELN zou kunnen gebruiken om naar data te verwijzen. Dit is specifiek voor mensen die een labnotebook gebruiken en dat verklaart het lage aantal geïnterviewden dat er over spreekt. Dit is een reële verwantschap.
- vlg naamconventie en vlg folderstructuur: specifiek aangegeven hoe een vaste folderstructuur wordt gebruikt in combinatie met een datum in de filenaam om volgbaarheid te bevorderen. Dit is een reële verwantschap.
- archief wat en archief waar: Wat wordt gearchiveerd lijkt wel verwantschap te hebben met waar wordt gearchiveerd. Dit komt met name door de verplichting in het beleid om in het 4TU datacenter op te slaan en de gewoonte van één van de onderzoekers om data vooral zoveel en zolang mogelijk op het rekencluster te laten staan. Is geen specifiek gebruik en er zijn maar twee bronnen, wat deze verwantschap onvoldoende sterk maakt om te laten bestaan.

Uit TABEL 106:

- rd eln en dd data verzamelen: twee voorbeelden waarbij verzamelde data in een labbook terechtkomen, het één van papier en het ander elektronisch. Dit is specifiek voor mensen die een labnotebook gebruiken en dat verklaart het lage aantal geïnterviewden dat er over spreekt. Dit is een reële verwantschap.
- rd gevoelig en bvlg contract: specifiek voor het geval dat voor bepaalde onderzoeksdata een ‘geheimhoudings’contract getekend moet worden. Dit is een reële verwantschap.
- dd link ont eln: specifiek voor ELN, wordt maar door enkele gebruikers van een ELN genoemd, waarbij is gevraagd of men het handig zou vinden als je vanuit je ELN kunt linken naar je data. Dit is specifiek voor mensen die een labnotebook gebruiken en dat verklaart het lage aantal geïnterviewden dat er over spreekt. Dit is een reële verwantschap.
- vind naamconventie en vlg naamconventie: Een naamconventie die helpt om te vinden, er staat bijvoorbeeld een duidelijke omschrijving in de naam en die helpt om te volgen doordat men bijvoorbeeld een datum meegeeft. Als je een naamconventie gebruikt, kan het zo zijn dat deze zo is ingericht dat het zowel de vindbaarheid als de volgbaarheid bevordert. Dat komt kennelijk weinig naar voren in de interviews, maar is wel een reële verwantschap.

Dit betekent dat verwantschappen inhoudelijk dus niet minder waardevol zijn als ze gebaseerd zijn op een kleiner aantal interviews (uit de bovenstaande voorbeelden, valt er maar één echt af).

Verwantschappen binnen de categorieën				
	Code 1	Code 2	Coëfficiënt	#Geïnterviewden
1	rd definitie	rd code	0.28	7
2	rd herkomst	rd meetdata	0.27	9
3	md definitie	md ja	0.40	7
4	md fdst persoonlijk	md nmc persoonlijk	0.29	6
5	md fdst groep	md nmc groep	0.24	4
6	md ja	md vorm	0.23	6
7	dd data bewerken	dd inhoud	0.29	4
8	dd data bewerken	dd eln	0.26	4
9	dd eln	dd inhoud	0.25	5
10	dd eln	dd link	0.21	3
11	del intern	del extern	0.36	4
12	del centraal	del sync and share	0.23	7
13	Bvlg authenticatie	Bvlg autorisatei	0.47	7
14	vind ja	vind folderstructuur	0.38	8
15	vind ja	vind notebook	0.26	5
16	vlg ja	vlg transformatie	0.49	9
17	vlg versie	vlg code	0.31	6
18	vlg ja	vlg code	0.22	9
19	vlg naamconventie	vlg folderstructuur	0.21	2
20	vlg ja	vlg versie	0.20	9
21	vlg transformatie	vlg versie	0.20	5
22	archief ja	archief eisen	0.50	6
23	archief ja	archief waar	0.46	7
24	archief eisen	archief hoe	0.40	5
25	archief eisen	archief waar	0.31	6
26	archief ja	archief hoe	0.30	5
27	archief wat	archief waar	0.26	2
28	archief eisen	archief wat	0.25	4

Tabel 105, Verwantschappen binnen de categorieën

Verwantschappen buiten de categorieën				
	Code 1	Code 2	Coëfficiënt	#Geïnterviewden
1	rd eln	dd dat bewerken	0.40	2
2	rd eln	dd data verzamelen	0.29	2
3	rd eln	dd inhoud	0.25	4
4	rd gevoelig	bvlg contract	0.25	3
5	md fdst groep	vind folderstructuur	0.27	5
6	md fdst persoonlijk	vind folderstructuur	0.21	4
7	md nmc persoonlijk	vlg naamconventie	0.22	6
8	dd link	ont eln	0.27	3
9	dd link	vind notebook	0.33	4
10	del code	opsl code	0.35	5
11	vind naamconventie	vlg naamconventie	0.25	3

Tabel 106, Verwantschappen buiten de categorieën

13.2. Conclusie co-occurrence/verwantschap

Verwantschap tussen codes geeft aan dat de codes mogelijk invloed op elkaar hebben. Hoe hoger de waarde van de coëfficiënt, des te sterker deze verwantschap waarschijnlijk is. Er is geen algemene wetmatigheid voor bij welke waarde van de coëfficiënt wordt gesproken van een sterke of zwakke verwantschap. Uit de analyse van de coëfficiënten in dit onderzoek blijkt dat bij een coëfficiënt van 0.3 of meer zeer waarschijnlijk van een bestaande verwantschap kan worden gesproken en dat bij coëfficiënten tussen de 0.2 en 0.3 verwantschap er waarschijnlijk is, maar waarschijnlijk ook zwak is.

14. Bijlage WorkSystem Snapshot

In de onderstaande paragrafen worden de bevindingen uit de literatuur die van toepassingen waren op het WSS naast relevante bevindingen uit de Interviews gelegd. De resultaten worden gebruikt voor een nieuw WSS.

14.1. Products and Services

In de eerste plaats wordt er gekken naar de bevindingen uit de literatuur en hoe deze terugkomen in de interviews. Waar nodig worden deze products and services aangepast (TABEL 107). Vervolgens worden de bevindingen uit de literatuur nagelopen op extra products and services (TABEL 108). Alle bevindingen worden vervolgens samengevoegd en gecontroleerd op vindbaarheid, volgbaarheid en langdurige opslag (TABEL 109). De uiteindelijke resultaten staan opgesomd in (TABEL 110).

Products and services na empirisch onderzoek				
#	**Oorspronkelijk product/Dienst**	**Uitleg na empirisch onderzoek**	**Nieuw product/dienst**	**Bron**
1.	Het werksysteem verzekert dat er gedurende het onderzoek voldoende betrouwbare opslagruimte beschikbaar is om de data op te slaan	Gedurende het onderzoek: Dit wordt het lopende onderzoek genoemd in de probleemstelling. Data: dit zijn de relevante data om het onderzoek (of een deel ervan) te kunnen herhalen, in de praktijk onderzoeksdata, metadata en datadocumentatie. De bevindingen uit de interviews geven geen reden de woorden voldoende en betrouwbaar aan te passen.	Het werksysteem verzekert dat er in het lopende onderzoek voldoende betrouwbare opslagruimte beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	RD MD DD
2.	Het werksysteem levert sync and share functionaliteit (‘academic dropbox’)	Er wordt veel gebruik gemaakt van commerciële share and sync diensten als Dropbox. Men maakt ook veel gebruik van project drive (een klassieke share oplossing met de mogelijkheid om de share ook open te zetten voor derden). Project drive (een eigen oplossing van de TU) biedt geen sync functionaliteit. De huidige bevinding suggereert een noodzaak die er in de praktijk niet is. Wel is het een wens van een deel van de geïnterviewden om project drive uit te breiden met sync functionaliteit (combinatie project drive en Surfdrive). Bij voorkeur wordt toegevoegd.	Het werksysteem levert functionaliteit om data te delen. Het werksysteem levert bij voorkeur functionaliteit om data te synchroniseren tussen externe datahouders en een centrale opslag.	DO DD
3.	Het werksysteem biedt de mogelijkheid dat data van electronic lab notebooks (ELN) verplaatst worden naar centrale active data opslag	Deze bevinding komt niet specifiek terug in de interviews. Geïnterviewde 12 zegt: ‘So I like the fact that I can run the experiment and put all the notes in my ELN here and then I go upstairs and because it's cloud based …..’ en ‘And [] the e-lab journals quite nice it's going up on the iPad as well so wherever as long as I can access a browser I can access my research notes which is really handy…. ‘ Functioneel betekent dit dus meer dat de data in de ELN vanaf elke locatie beschikbaar moeten zijn.	Het werksysteem biedt de mogelijkheid een ELN te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.	Zie uitleg
4.	Het werksysteem maakt het mogelijk om de data van elk device (pc’s, laptops, tablets, telefoons, laboratoriuminstrumenten enz.) te kopiëren.	De benoemde devices in de bevinding zijn vormen van externe datahouders. Eerder is synchroniseren gedefinieerd als: Data naar een tweede (of meer dan dat) locatie repliceren. Dat kan een eenmalige actie zijn (een kopie), het kan realtime zijn, waarbij de beide omgevingen voortdurend gelijk zijn (spiegelen) of dat data op gezette tijden wordt gerepliceerd.	Het werksysteem levert functionaliteit om data te synchroniseren tussen externe datahouders en een centrale opslag.	LIT

#	Oorspronkelijk product/Dienst	Uitleg na empirisch onderzoek	Nieuw product/dienst	Bron
		Met andere woorden, een kopie is een speciale vorm van synchronisatie.		
5.	Het werksysteem biedt de mogelijkheid om data naar meerdere locaties/systemen te distribueren/repliceren.	Vergelijkbaar met hierboven, maar nu in de andere richting. Hoewel synchroniseren zich niet beperkt tot één richting, wordt het er expliciet bijgezet.	Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen externe datahouders en een centrale opslag. (uiteindelijke versie)	LIT
6.	Het werksysteem biedt de mogelijkheid een eigen naamgevingformaat en (mappen)structuur voor de data aan te houden.	Er wordt onderscheid gemaakt tussen de mogelijkheid om persoonlijke invulling te geven aan folderstructuren en naamconventies en om een groepsinvulling te geven. In dat laatste geval geven de geïnterviewden 1, 8 en 10 aan dat ze het opleggen, waarbij de geïnterviewden van 8 daarin ook strikt handhaven. In de andere gevallen wordt het aan de individuele onderzoeker overgelaten of wordt het geadviseerd vanuit de PI. Een aantal van de geïnterviewden geeft aan dat het goed zou zijn om templates te kunnen gebruiken (in het algemeen, niet alleen voor folderstructuren en naamconventies).	Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek aan te houden waarbij formats in templates worden beschreven.	MD ON WW
7.	Het werksysteem verzekert dat onderzoeksdata vindbaar en begrijpelijk zijn door de data in combinatie met de bijbehorende metadata en andere documentatie op te slaan	Dit wordt omgeschreven naar de begrippen zoals geformuleerd na de interviews.	Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.	RD MD DD
8.	Het werksysteem verzekert dat backupmogelijkheden gedurende het onderzoek beschikbaar zijn	Een aantal geïnterviewden gebruikt een share and sync oplossing bij wijze van backup. De geïnterviewde 1 geeft aan dat dit eerder is gebruikt voor een restore (deze geïnterviewde is overigens de enige die aangeeft ooit een restore te hebben moeten doen). Veel geïnterviewden geven aan een periodieke kopie naar een externe datahouder te maken. Beide punten worden afgedekt met de bevinding dat het werksysteem functionaliteit levert om data te synchroniseren tussen externe datahouders en een centrale opslag. Een zestal geeft aan de data op te slaan op TU systemen, waarvan men verwacht dat ze door de ICT afdeling worden geback-upped.	Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.	DB
9.	Het werksysteem zorgt dat het aanmaken van metadata bij voorkeur geautomatiseerd plaatsvindt	Geïnterviewden 3, 8 en 12 geven aan dat ze deze mogelijkheid hebben. Geïnterviewde 7 wil het graag hebben. De voorkeur is misschien daardoor te sterk, maar de mogelijkheid zou er moeten zijn. Ook hier zou de mogelijkheid om templates te gebruiken een aanvulling kunnen zijn.	Het werksysteem biedt de mogelijkheid om geautomatiseerd metadata aan te maken, bij voorkeur m.b.v. een template.	MD
10.	Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld	Interactie tussen het werksysteem en andere systemen is in de interviews niet naar voren gekomen, dat klinkt meer als een eventuele wens. Het werken in een	Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een	Niet specifiek

#	Oorspronkelijk product/Dienst	Uitleg na empirisch onderzoek	Nieuw product/dienst	Bron
	een userinterface) en voor machines (bijvoorbeeld een API). Dit is noodzakelijk om met het werksysteem te kunnen werken.	ELN, het aanmaken van files en folders en andere soortgelijke activiteiten wel. Daar zijn geen specifieke eisen en wensen bij naar voren gekomen.	userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).	
11.	Het werksysteem levert een mechanisme om versiebeheer op de mappen, bestanden en data te kunnen toepassen	Versiebeheer komt veelvuldig terug in de interviews, zowel op documenten als op data (met name code). Het is een mechanisme om volgbaarheid te vergroten (zie ook Volgbaarheid versies). Men houdt versie geautomatiseerd bij in share and sync software, in code repositories (Git Hub. Git Lab en Bit Bucket) en soms in geautomatiseerde metadata. Sommigen geven met de hand een volgnummer mee aan een document. Volgnummers op mappen, worden niet genoemd. Met name geautomatiseerd versiebeheer, lijkt goed te passen bij de voorkeurswerkwijze van de meeste wetenschappers.	Het werksysteem levert een mechanisme om gautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.	MD DO WW
12.	Het werksysteem levert de mogelijkheid om data te vernietigen	In combinatie met het loggen van de activiteit, zou het vernietigen van data volgbaar zijn.	Het werksysteem maakt het mogelijk data te vernietigen en logt deze activiteit.	RD Logging
13.	Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan	Dit komt expliciet en impliciet niet terug in de interviews. Het lijkt echter voor de hand te liggen dat men geen data wil verliezen. De bevinding wordt ongewijzigd gehandhaafd.	Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.	LIT
14.	Het werksysteem levert de mogelijkheid om data te transformeren om aan regelingen te voldoen (bijvoorbeeld anonimisering)	Transformeren van data om aan regelingen te voldoen, komt expliciet en impliciet niet terug in de interviews. Get wordt wel benoemd als onderdeel van beveiliging. Het wordt als zodanig opgenomen.	Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek te beveiligen.	DB
15.	Het werksysteem kan unieke identifiers aanmaken voor een dataset (Persistent Identifiers, PID).	Het archiveren bij een externe, daartoe gespecialiseerde repository, is buiten scope geplaatst.	Deze valt af op basis van de bevindingen uit architectuur	ARC
16.	Het werksysteem verzekert dat de data selectief beschikbaar gemaakt kunnen worden aan anderen	Selectief beschikbaar stellen is een combinatie van data delen, authenticatie (met wie wordt gedeeld)en autorisatie (wat de persoon met wie gedeeld wordt, mag). Dit betekent dat er aanvullend op de eerdere dienst om data delen mogelijk te maken, een extra dienst moet komen, die het mogelijk maakt relevante data te beveiligen. Het ligt hierbij voor de hand dat data kunnen versleutelen, anonimiseren, pseudonimiseren daar ook bij hoort en dat dus beveiliging en niet alleen autorisatie en authenticatie genoemd hoeven te worden.	Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek te beveiligen.	DB DD
17.	Het werksysteem moet dermate flexibel zijn, dat het een grote variëteit aan gebruik kan ondersteunen	Dit wordt aangepast op basis van bijlage ontwerp.	Het werksysteem biedt de onderzoeker voldoende flexibiliteit om het per situatie aan te passen	ONT

#	Oorspronkelijk product/Dienst	Uitleg na empirisch onderzoek	Nieuw product/dienst	Bron
18.	Het werksysteem draagt er zorg voor dat actieve onderzoeksdata snel toegankelijk zijn. Het kan sterke rekenkracht (computational resources) nodig hebben	De stelling wordt herschreven met onderzoeksdata uit het lopende onderzoek.	Het werksysteem draagt er zorg voor dat onderzoeksdata uit het lopende onderzoek snel toegankelijk zijn. Het kan sterke rekenkracht (computational resources) nodig hebben	RD
19.	Het werksysteem ondersteunt langdurig lopende experimenten	Dit wordt niet expliciet benoemd in de interviews, geen verandering nodig.	Het werksysteem ondersteunt langdurig lopende experimenten	LIT
20.	Het werksysteem moet privacyeisen, copyrighteisen en eisen uit wet- en regelgeving kunnen waarborgen	Dit valt in de interviews onder het begrip beveiliging. En wordt benoemd in bevinding 16	Volgt bevinding 16	16
21.	Het werksysteem biedt opslag voor de RDM infrastructuur die gelaagd moet zijn naar prijs en daarbij behorende veiligheid van de data (goedkoop en relatief onveilig voor reproduceerbare data en duur en veilig voor niet reproduceerbare data)	Dit komt niet terug in de bevindingen van de interviews en blijft dus ongewijzigd staan.	Het werksysteem biedt opslag voor de RDM infrastructuur die gelaagd moet zijn naar prijs en daarbij behorende veiligheid van de data (goedkoop en relatief onveilig voor reproduceerbare data en duur en veilig voor niet reproduceerbare data)	LIT

Bronnen
LIT: THEORETISCH KADER
RD: ONDERZOEKSDATA
MD: METADATA
DD: DATADOCUMENTATIE
DO: DATA OPSLAG
DB: DATA BEVEILIGING
ON: ONTWERP
WW: CONCLUSIE DATA BEVEILIGING
ARC: ARCHIVEERBAARHEID

Tabel 107, Products and services na empirisch onderzoek

In de tabel hieronder staan overige products and services die uit de interviews naar voren zijn gekomen. De bron staat er steeds bij. Sommige worden afgedekt door de bevindingen uit één van de vorige twee tabellen. Dat staat dan benoemd. In het andere geval worden de bevindingen één op één overgenomen.

Products and services uit de interviews			
#	**Oorspronkelijk product/Dienst**	**Bron**	**Nieuw product/dienst**
1.	Het werksysteem biedt de mogelijkheid om templates te gebruiken (o.a. voor folderstructuren, naamconventies, metadata)	Dit komt expliciet uit de bijlage ontwerp	Dit wordt afgedekt door de aangepaste bevinding 6.
2.	Het werksysteem biedt de onderzoeker voldoende flexibiliteit om het per situatie aan te passen	Dit komt expliciet uit de bijlage ontwerp	Dit wordt afgedekt door de aangepaste bevinding 17.
3.	Het werksysteem neemt de onderzoeker ondersteunend werk uit handen	Dit komt expliciet uit de bijlage ontwerp	Het werksysteem neemt de onderzoeker ondersteunend werk uit handen

4.	Het werksysteem biedt share and sync opslag met minimaal 8TB opslag per onderzoek	Dit komt expliciet uit de bijlage ontwerp, minimaal 8TB wordt toegevoegd aan 1 en vervangt 'voldoende'. Deze bevinding vervangt dan die van 1.	Het werksysteem verzekert dat er in het lopende onderzoek minimaal 8TB betrouwbare opslagruimte beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.
5.	Het werksysteem biedt de mogelijkheid data van externe datahouders te synchroniseren met de 8TB opslag	Dit komt expliciet uit de bijlage ontwerp	Dit wordt afgedekt door de aangepaste bevinding 5
6.	Het werksysteem biedt de mogelijkheid een ELN te gebruiken	Dit komt expliciet uit de bijlage ontwerp	Dit staat in bevinding 3
7.	Het ELN in het werksysteem biedt de mogelijkheid om o.a. activiteiten en keuzes te loggen	Dit komt expliciet uit de bijlage ontwerp	Het ELN in het werksysteem biedt de mogelijkheid om o.a. activiteiten en keuzes te loggen
8.	Het ELN in het werksysteem biedt de mogelijkheid om te linken naar data	Dit komt expliciet uit de bijlage ontwerp	Het ELN in het werksysteem biedt de mogelijkheid om te linken naar data
9.	Het ELN in het werksysteem kan specifiek gemaakt worden voor bepaalde vakgebieden	Dit komt expliciet uit de bijlage ontwerp	Het ELN in het werksysteem kan specifiek gemaakt worden voor bepaalde vakgebieden
10.	Het werksysteem biedt een opslagplaats voor langdurige relevante data die voldoende snel (sneller dan tape) opgehaald kan worden	Dit komt expliciet uit de bijlage ontwerp	Het werksysteem biedt een opslagplaats voor langdurige relevante data die voldoende snel (sneller dan tape) opgehaald kan worden

Tabel 108, Products and services uit de interviews

De resulterende bevindingen uit de voorgaande twee tabellen worden samengevoegd en beoordeeld op vindbaarheid, volgbaarheid en langdurige opslag.

Products and services en vindbaarheid, volgbaarheid en langdurige opslag					
#	**Product/dienst**	**VDB**	**VLG**	**LO**	
1.	Het werksysteem verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	X			Bekende locatie maakt vinden makkelijker.
2.	Het werksysteem levert functionaliteit om data te delen.				Delen is een sterke wens, maar draagt niet bij aan vdb of vgb of lo.
3.	Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen externe datahouders en een centrale opslag.	X			Data staan op de centrale opslag na synchronisatie. Dat verhoogt vdb.
4.	Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek aan te houden waarbij formats in templates worden beschreven.	X	X		Relevante data en folderstructuren en naamconventies tezamen vormen een goede basis voor vdb en vgb.
5.	Het werksysteem biedt de mogelijkheid een ELN te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.	X	X		ELN kan voor vdb en vlg zorgen.
6.	Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.	X	X		Vdb en vgb zijn onderdeel van de stelling.
7.	Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.

#	Product/dienst	VDB	VLG	LO	
8.	Het werksysteem biedt de mogelijkheid om geautomatiseerd metadata aan te maken, bij voorkeur m.b.v. een template.	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
9.	Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).	X	X		Om data te kunnen vinden en volgen is een interface om mee te zoeken noodzakelijk.
10	Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.		X		Helpt datatransformaties (vlg) inzichtelijk te maken.
11	Het werksysteem maakt het mogelijk data te vernietigen en logt deze activiteit.	X	X		De laatste bewerking op data. Door de logging weet je dat de data verdwenen zijn en niet gevonden kunnen worden
12	Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.		X		Op voorwaarde van geschikte logging volgbaar.
13	Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek te beveiligen.		X		Levert bij goede logging volgbaarheid op omdat inzichtelijk is wie welke bewerking met data heeft uitgevoerd.
14	Het werksysteem biedt de onderzoeker voldoende flexibiliteit om het per situatie aan te passen				Levert geen directe bijdrage aan vdb en vgb en lo.
15	Het werksysteem neemt de onderzoeker ondersteunend werk uit handen				Levert geen directe bijdrage aan vdb en vgb en lo.
16	Het ELN in het werksysteem biedt de mogelijkheid om o.a. activiteiten en keuzes te loggen	X	X		Helpt vdb en vgb als je bijv. logt waar je iets hebt opgeslagen of wat je met een dataset hebt gedaan.
17	Het ELN in het werksysteem biedt de mogelijkheid om te linken naar relevante data in het lopende onderzoek.	X	X		Vgb want linken naar data die transformaties beschrijven.
18	Het ELN in het werksysteem kan specifiek gemaakt worden voor bepaalde vakgebieden		X		Bijvoorbeeld een procesbeschrijving in chemische wetenschappen.
19	Het werksysteem biedt een opslagplaats voor koude relevante data die voldoende snel (sneller dan tape) opgehaald kan worden			X	Geen data in lopend onderzoek.

VDB: Geeft antwoord op de vraag: Waar zijn de relevante data uit het lopende onderzoek (fysiek en/of logisch)?
VLG: Geeft antwoord op de vraag: Kan ik de wijzigingen die de relevante onderzoeksdata ondergaan volgen op basis van de beschikbare informatie?
LO: Geeft antwoord op de vraag Kan ik relevante data uit het lopende onderzoek ergens langdurig opslaan?

Tabel 109, Products and services en vindbaarheid, volgbaarheid en langdurige opslag

Als de products and services waarvoor geldt dat ze niet bijdragen aan vindbaarheid, volgbaarheid en/of langdurige opslag worden verwijderd, dan blijft TABEL 110 over.

Products and services, uiteindelijke lijst					
#	**Proct/Dienst**	**VDB**	**VLG**	**LO**	
17.	Het werksysteem verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	X			Bekende locatie maakt vinden makkelijker.
18.	Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen externe datahouders en een centrale opslag.	X			Data staan op de centrale opslag na synchronisatie. Dat verhoogt vdb.
19.	Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek aan te houden waarbij formats in templates worden beschreven.	X	X		Relevante data en folderstructuren en naamconventies tezamen vormen een goede basis voor vdb en vgb.
20.	Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.	X	X		Vdb en vgb zijn onderdeel van de stelling.
21.	Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.
22.	Het werksysteem biedt de mogelijkheid om geautomatiseerd metadata aan te maken, bij voorkeur m.b.v. een template.	X	X		Metadata helpen bij vindbaarheid en volgbaarheid.
23.	Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).	X	X		Om data te kunnen vinden en volgen is een interface om mee te zoeken noodzakelijk.
24.	Het werksysteem maakt het mogelijk data te vernietigen en logt deze bewerking.	X	X		De laatste bewerking op data. Door de logging weet je dat de data verdwenen zijn en niet gevonden kunnen worden
25.	Het werksysteem biedt de mogelijkheid een ELN te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.	X	X		ELN kan voor vdb en vlg zorgen.
26.	Het ELN in het werksysteem biedt de mogelijkheid om o.a. activiteiten en keuzes te loggen	X	X		Helpt vdb en vgb als je bijv. logt waar je iets hebt opgeslagen of wat je met een dataset hebt gedaan.
27.	Het ELN in het werksysteem biedt de mogelijkheid om te linken naar relevante data in het lopende onderzoek.	X	X		Vgb want linken naar data die transformaties beschrijven.
28.	Het ELN in het werksysteem kan specifiek gemaakt worden voor bepaalde vakgebieden		X		Bijvoorbeeld een procesbeschrijving in chemische wetenschappen.
29.	Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek te beveiligen.		X		Het verschaffen van autorisaties, i.c.m. logging bevordert de volgbaarheid.
30.	Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan. Databewerkingen worden gelogd.		X		Op voorwaarde van geschikte logging volgbaar.

#	Proct/Dienst	VDB	VLG	LO	
31.	Het werksysteem levert een mechanisme om gautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.		X		Helpt datatransformaties inzichtelijk te maken.
32.	Het werksysteem biedt een opslagplaats voor koude relevante data die voldoende snel (sneller dan tape) opgehaald kan worden			X	Langdurige opslag van data in lopend onderzoek en daarna.
VDB: Geeft antwoord op de vraag: Waar zijn de relevante data uit het lopende onderzoek (fysiek en/of logisch)? VLG: Geeft antwoord op de vraag: Kan ik de wijzigingen die de relevante onderzoeksdata ondergaan volgen op basis van de beschikbare informatie? LO: Geeft antwoord op de vraag Kan ik relevante data uit het lopende onderzoek ergens langdurig opslaan?					

Tabel 110, Products and services, uiteindelijke lijst.

14.2. Activiteiten

Activiteiten werden eerder afgeleid van de in de literatuur gevonden products and services. Ze kunnen nu bepaald worden op basis van de uitkomsten van de interviews (en zijn dan per definitie specifiek voor de case organisatie). Ze staan in TABEL 111, inclusief de bron. De vindbaarheid, volgbaarheid en langdurige opslag wordt ook benoemd. Vervolgens wordt gecontroleerd of de bevindingen uit de literatuur zijn afgedekt (TABEL 112). Zo niet dan worden ze toegevoegd als ze niet strijdig zijn met de interviews en worden alleen degenen die een bijdrage leveren aan vindbaarheid, volgbaarheid en langdurige opslag opgenomen (TABEL 113).

Work practices op basis van de interviews						
#	**Activiteit**	**VDB**	**VLG**	**LO**	**Uitleg**	**Bron**
1.	De onderzoeker herhaalt (delen van) het onderzoek	X	X	X	Hiervoor is vdb, vgb en evt. lo nodig	RD
2.	De onderzoeker bewerkt onderzoeksdata op 3 verschillende niveaus	X	X		Om te kunnen bewerken, moet data vindbaar zijn. Voor verschillende niveaus volgbaar	RD
3.	De onderzoeker genereert/bewerkt verschillende soorten data (meetdata, simulatiedata, et cetera)				Niet specifiek vdb of vgb of lo.	RD
4.	De onderzoeker ruimt data op	X			Als je weet dat het is opgeruimd vdb	RD
5.	De onderzoeker verzamelt data die van binnen of buiten de TU Komt				Niet specifiek vdb of vgb of lo.	RD
6.	De onderzoeker maakt metadata aan/gebruikt metadata	X	X		Metadata kunnen helpen bij vdb en vgb	MD
7.	De onderzoeker genereert geautomatiseerd metadata	X	X		Metadata kunnen helpen bij vdb en vgb	MD
8.	De onderzoeker beschrijft metadata in lablogboeken	X	X		Metadata kunnen helpen bij vdb en vgb	MD
9.	De PI adviseert een folderstructuur of legt deze op	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
10.	De PI adviseert een naamconventie of legt deze op	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
11.	De onderzoeker gebruikt een folderstructuur	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
12.	De onderzoeker gebruikt een naamconventie	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
13.	De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verzameld		X		Hoe het wordt verzameld, kan iets zeggen over vgb	DD
14.	De onderzoeker houdt georganiseerd aantekeningen bij over de resulterende databestanden (wat ze zijn, waar ze staan e.d.)	X	X		Vdb staat in de stelling. weten hoe elk databestand is bewerkt is vgb	DD
15.	De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verwerkt		X		weten hoe elk databestand is bewerkt is vgb	DD
16.	De onderzoeker verwijst naar een resulterend databestand met een link in de datadocumentatie	X			Beschrijft vdb	DD

#	Activiteit	VDB	VLG	LO	Uitleg	Bron
17.	De onderzoeker slaat relevante data voor het lopende onderzoek op diverse locaties op				Er staat niet bij hoe dit vdb of vgb is of lo.	DO
18.	Er worden backups gemaakt van relevante data voor het lopende onderzoek	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.	DO
19.	De onderzoeker deelt relevante data uit het lopende onderzoek				Zegt niets over vdb en vgb	DD
20.	De onderzoeker beveiligt de relevante data voor het lopende onderzoek				Zegt niets over vdb en vgb	DB
21.	De relevante data bewerkingen en benaderingen worden gelogd		X		Zo kun je volgen wat er met de data gebeurt	DB
22.	De onderzoeker pas versiebeheer toe		X		Versiebeheer helpt bij vgb	VLG
23.	De onderzoeker slaat voldoende relevante data uit het lopende onderzoek op op een plek waar hij/zij het kan terugvinden	X			Vdb het staat er letterlijk	ARC

VDB: Geeft antwoord op de vraag: Waar zijn de relevante data uit het lopende onderzoek (fysiek en/of logisch)?
VLG: Geeft antwoord op de vraag: Kan ik de wijzigingen die de relevante onderzoeksdata ondergaan volgen op basis van de beschikbare informatie?
LO: Geeft antwoord op de vraag Kan ik relevante data uit het lopende onderzoek ergens langdurig opslaan?

RD: ONDERZOEKSDATA
MD: METADATA
DD: DATADOCUMENTATIE
DO: DATA OPSLAG
DB: DATA BEVEILIGING
VLG: VOLGBAARHEID
ARC: ARCHIVEERBAARHEID

Tabel 111, Work practices op basis van de interviews

Activiteiten uit de literatuur	
Activiteit uit literatuur	Overnemen na empirisch onderzoek?
Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan	Staat niet in TABEL 111 en is niet strijdig met de interviews
Het werksysteem keurt de aanvraag en kent ruimte toe of niet;	Staat niet in TABEL 111 en is niet strijdig met de interviews
Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte.	Staat niet in TABEL 111 en is niet strijdig met de interviews
De researcher verplaatst (meta)data op de geboden opslagruimte;	Staat niet in TABEL 111 en is niet strijdig met de interviews
De researcher verwijdert data van de geboden opslagruimte;	Genoemd in de interviews
De researcher bedenkt en gebruikt een heldere mappenstructuur op het werksysteem;	Genoemd in de interviews
Het werksysteem synchroniseert lokale onderzoeksdata en metadata met de centraal aangeboden opslagruimte;	Benoemd bij products and services
De researcher slaat zijn (meta)data op op de geboden opslagruimte;	Genoemd in de interviews
De researcher wijzigt (meta)data op de geboden opslagruimte;	Genoemd in de interviews
Het werksysteem voert versiebeheer uit over de (meta)data;	Benoemd bij products and services
Het werksysteem levert de researcher mogelijkheden voor de aanmaak van metadata;	Benoemd bij products and services
Het werksysteem voert backups uit op basis van de eisen van de researcher;	Benoemd bij products and services

Activiteit uit literatuur	Overnemen na empirisch onderzoek?
De researcher vraagt een PID aan bij het werksysteem voor een relevante datasets en bijbehorende metadata;	Niet meer relevant nadat archiveren langdurig opslaan is gemaakt
Het werksysteem maakt een PID aan voor een dataset en bijbehorende metadata.	Niet meer relevant nadat archiveren langdurig opslaan is gemaakt

Tabel 112, Work practices uit de literatuur

De voorgaande twee tabellen resulteren in TABEL 113.

Activiteiten, uiteindelijke lijst						
#	**Activiteit**	**VDB**	**VLG**	**LO**		
23.	Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan	X				LIT
24.	Het werksysteem keurt de aanvraag en kent ruimte toe of niet	X				LIT
25.	Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte	X				LIT
26.	De researcher verplaatst (meta)data op de geboden opslagruimte	X				LIT
27.	De onderzoeker ruimt data op	X			Als je weet dat het is opgeruimd vdb	RD
28.	De onderzoeker verwijst naar een resulterend databestand met een link in de datadocumentatie	X			Beschrijft vdb	DD
29.	De onderzoeker slaat voldoende relevante data uit het lopende onderzoek op op een plek waar hij/zij het kan terugvinden	X			Vdb het staat er letterlijk	ARC
30.	De onderzoeker bewerkt onderzoeksdata op 3 verschillende niveaus	X	X		Om te kunnen bewerken, moet data vindbaar zijn. Voor verschillende niveaus volgbaar	RD
31.	De onderzoeker maakt metadata aan/gebruikt metadata	X	X		Metadata kunnen helpen bij vdb en vgb	MD
32.	De onderzoeker genereert geautomatiseerd metadata	X	x		Metadata kunnen helpen bij vdb en vgb	MD
33.	De onderzoeker beschrijft metadata in lablogboeken	X	X		Metadata kunnen helpen bij vdb en vgb	MD
34.	De PI adviseert een folderstructuur of legt deze op	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
35.	De PI adviseert een naamconventie of legt deze op	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
36.	De onderzoeker gebruikt een folderstructuur	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD
37.	De onderzoeker gebruikt een naamconventie	X	X		Folderstructuren en naamconventies kunnen helpen bij vdb en vgb	MD

#	Activiteit	VDB	VLG	LO		
38.	De onderzoeker houdt georganiseerd aantekeningen bij over de resulterende databestanden (wat ze zijn, waar ze staan e.d.)	X	X		Vdb staat in de stelling. weten hoe elk databestand is bewerkt is vgb	DD
39.	Er worden backups gemaakt van relevante data voor het lopende onderzoek	X	X		Backups geven tussentijdse versie en helpen bij vind- en volgbaarheid.	DO
40.	De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verzameld		X		Hoe het wordt verzameld, kan iets zeggen over vgb	DD
41.	De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verwerkt		X		weten hoe elk databestand is bewerkt is vgb	DD
42.	De relevante data bewerkingen en benaderingen worden gelogd		X		Zo kun je volgen wat er met de data gebeurt	DB
43.	De onderzoeker pas versiebeheer toe		X		Versiebeheer helpt bij vgb	VLG
44.	De onderzoeker herhaalt (delen van) het onderzoek	X	X	X	Hiervoor is vdb, vgb en evt. lo nodig	RD

LIT: THEORETISCH KADER
RD: ONDERZOEKSDATA
DD: DATADOCUMENTATIE
ARC: ARCHIVEERBAARHEID
MD: METADATA
DO: DATA OPSLAG
DB: DATA BEVEILIGING
VLG: VOLGBAARHEID

Tabel 113, Work practices, uiteindelijke lijst

14.3. Technology

Technology bestaat met name uit hulpmiddelen die het werk ondersteunen. Een deel ervan is algemeen (general purpose) en een deel ervan is specifiek voor het werk (tailored). In de interviews worden voor technology genoemd:

- De diverse toegepaste opslagsystemen (zoals benoemd in DATA OPSLAG CONCLUSIE);

Data delen vanaf de verschillende toegepaste opslagsystemen (zoals benoemd in DATA DELEN CONCLUSIEHet lijkt er niet op dat de co-occurrence extra informatie oplevert t.o.v. het voorgaande.

- Data delen conclusie);
- Data beveiligingstechnieken (zoals benoemd in CONCLUSIE DATA BEVEILIGING);
- Versiebeheer geautomatiseerd uitgevoerd;
- ELN's (en ook papieren logboeken).

Dropbox is een algemene share and sync tool die wordt ingezet voor data opslag, data delen versiebeheer en datasynchronisatie van relevante data.

14.4. Information

Dit gaat met name om information die wordt uitgewisseld tussen mensen en systemen, mensen en mensen of systemen en systemen o het werk gedaan te krijgen. In de interviews worden genoemd:

- Adviezen van data stewards aan onderzoekers (inclusief het opmaken van een data management plan);
- Het data management plan, met de planning van het databeheer.
- Adviezen van senior researchers (meestal de PI) aan de onderzoekers (o.a. voor folderstructuren en naamconventies);
- Versiebeheer handmatig uitgevoerd;
- Ruimte aanvragen door een PI;
- Overzichten uit logging.

14.5. Participants

De participants voeren het werk uit (m.b.v. systemen). Ze zijn al eerder voor de case organisatie gespecificeerd in de bijlage SELECTIE GEÏNTERVIEWDEN:

- Principle Investigators (PI's);
- Onderzoekers;
- Data Stewards;
- IT ondersteuners (in de afdeling/sectie).

14.6. Customers

Kijkend naar de lijst met participants zijn dat met name de PI's en onderzoekers.

15. Bijlage expert review

Hieronder volgen de selectie van het expert panel, de uitvoering van de review en de correspondentie tijdens de review.

15.1. Bijlage Selectie expert panel

De belangrijkste componenten in het ontwikkelde WSS zijn products and services en technology/infrastructure. Het WSS moet goed aansluiten bij de praktijk onderzoekers. Ten aanzien van de bestaande praktijk, zijn de uitkomstem mogelijk innovaties. Hierom worden de volgende personen geselecteerd:

Een proces architect: Een ervaren Business Analist en Proces Architect (30 jaar ervaring met BPM, BA en Proces Architectuur).

Een infrastructuur architect: Ervaren ICT infrastructuur architect met ruime ervaring bij meerdere Nederlandse universiteiten. Ervaring met het maken van roadmaps en high-level designs voor ICT technische infrastructuur alsmede generieke applicaties t.b.v. die infrastructuur, persoonlijke – en groepsproductiviteit.

Een innovatie adviseur: Heeft zelf onderzoek gedaan, werkt afwisselend parttime binnen een onderzoeksafdeling van de verschillende faculteiten van de instelling en houdt zich verder bezig met (ICT) innovatie.

Alle drie hebben ze binding met ICT en met de gebruikers van ICT diensten in hun dagelijkse werk.

15.2. Bijlage Expert review

In deze bijlage wordt de expert review door drie verschillende experts beschreven. De expert review is gedaan volgens de Delphi methode. Drie experts kregen onafhankelijk van elkaar de products and services zoals beschreven na de interviews, uitleg over het proces hoe tot deze products and services gekomen was en wat extra uitleg over gebruikte begrippen. Vervolgens werd ze gevraagd elke product/dienst te scoren op een Likert schaal: kan ik me helemaal niet in vinden, kan ik me niet in vinden, sta ik neutraal in, kan ik me in vinden, kan ik me helemaal in vinden, geen oordeel. Het verschil tussen sta ik neutraal in en geen oordeel is:

- Sta ik neutraal in: Iemand is niet voor of tegen;
- Geen oordeel: iemand geeft aan dat hij/zij zich niet geschikt acht om hier een gewogen oordeel te geven.

Bovendien werd verzocht de scores van commentaar te voorzien. Aan de hand van dit commentaar werd extra informatie verschaft en werd de experts verzocht nogmaals te scoren en commentaar te geven. Ook hierna werd extra informatie gegeven en kwam het tot een eindscore in de derde ronde. Alles verliep schriftelijk en anoniem. De experts hebben elkaars namen nooit gezien en wisten niet van elkaar dat ze in de reviewgroep zaten.
Hieronder volgens na elkaar steeds een tabel met scores en enkele tabellen met commentaar. Per ronde wordt een korte conclusie getrokken over de scores en het commentaar.

15.2.1. Scores na de eerste ronde

Scores na de eerste ronde							
		Kan ik me helemaal niet in vinden	**Kan ik me niet in vinden**	**Sta ik neutraal in**	**Kan ik me in vinden**	**Kan ik me geheel in vinden**	**Geen oordeel**
	Opslag						
1	Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.		3		12		
2	Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.		3		2	1	
3	Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.		23		1		
4	Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel (sneller dan tape) opgehaald kunnen worden.		2		13		
	Relevante data						
5	Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek aan te houden waarbij formats in templates beschreven kunnen worden.		3	2		1	
6	Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.		3		2	1	
7	Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.		3		12		
8	Het *werksysteem* maakt het mogelijk data te vernietigen en logt deze activiteit.		3			2	1
9	Het *werksysteem* biedt de mogelijkheid om geautomatiseerd *metadata* aan te maken, bij voorkeur m.b.v. een template.				13		2
	Interactie						
10	Het *werksysteem* biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).		2			13	
	Beveiligen						
11	Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek te beveiligen.			23		1	
12	Het *werksysteem* verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.			2		13	
	ELN						
13	Het *werksysteem* biedt de mogelijkheid een *ELN* te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.		3	2		1	
14	Het *werksysteem* biedt een *ELN* aan dat de mogelijkheid heeft om o.a. work practices en keuzes te loggen.			2	3	1	
15	Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om te linken naar relevante data in het lopende onderzoek.				12	3	
16	Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden.		2			13	

Tabel 114, scores expert review ronde 1

15.2.2. Commentaar na de eerste ronde

Commentaar en reactie over de categorie opslag bij de eerste ronde					
		Commentaar 1 na eerste ronde	**Commentaar 2 na eerste ronde**	**Commentaar 3 na eerste ronde**	**Commentaar en voorstellen Bert Kuipers**
	Opslag				
1	Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	8Tb is voor veel onderzoeken te veel, maar voor sommige onderzoeken te weinig (denk aan lang lopende onderzoeken met videobeelden als bronmateriaal). Sommige data wordt gebruikt voor meerdere onderzoeken (denk aan scans van historische landkaarten). Er is dan behoefte aan opslag dat gebruikt kan worden voor meerdere onderzoeken, zonder onnodige kopieën te genereren.	Scoping van je werksysteem is verwarrend: je praat over lopende onderzoeken. Maar hoe zit het met afgesloten onderzoeken, waar een andere wetenschapper in een nieuw onderzoek het onderzoek wil gebruiken, of herhalen, of andere hypothese op basis van de data wil testen? Opslaan wijst niet alleen in het werksysteem op opslag (infrastructuur), maar ook functionaliteiten (en evt. mensen) die dit mogelijk maken. Een wat smalle beschrijving. Overig: 8TB voelt wat arbitrair en wordt niet onderbouwd. Op dit niveau lijkt me dit ook wat prematuur (implementatie). Je kan wel iets roepen over een 'afgebakende hoeveelheid opslag'.	Steven Alter definieert het werksysteem als een systeem waarin menselijke participants en/of machines werk (processen/activiteiten) uitvoeren waarbij gebruik gemaakt wordt van informatie, techniek en andere bronnen om te komen tot specifieke producten of diensten voor specifieke interne en/of externe customers. De vraag is of de term 'relevante data' op basis van bovenstaande beschrijving van Alter de lading voldoende dekt. Verder noem je een ondergrens voor de opslagcapaciteit behorende bij een werksysteem, is er ook een bovengrens of moet de oplossing oneindig schaalbaar zijn?	Voor scope en beschrijving wil ik graag verwijzen naar het bijgeleverde Word document. T.a.v. de 8TB lees ik dat iedereen dat te specifiek vindt. Mijn voorstel zou zijn om dat er uit te halen en te vervangen voor voldoende (door de wetenschapper te bepalen) opslagruimte. T.a.v. van nummer 3: wetenschappers zijn onderdeel van het werksysteem (participants en customers). T.a.v. nummer 2, ik kan niet specifieker worden dan de literatuur en interviews me aan ruimte geven. Ik kan wel anders formuleren.
2	Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.	Data analyseren in het vliegtuig onderweg naar congres is bij meerdere onderzoekers een wens.	Waarom in twee richtingen? Ik heb het gevoel dat er nog een onderscheid te maken is bij onderzoeksdatahouders: puur data-verzamelende datahouders (bijv. specifieke meetapparatuur) en lokale datahouders waar de pi's analyses doen op 1 of meerdere databronnen en evt. bewerkingen op doen. Klopt dat? Voor dat laatste type kan ik me twee richtingen voorstellen. Ook hier een scope-onduidelijkheid: is het DOEN van onderzoek in scope van het werksysteem? Het verzamelen en bewerken van data een onderdeel van de scope van het werksysteem? Of alleen het vastleggen en documenteren? En het vastleggen: is de decentrale opslag ook in scope van dit werksysteem, of alleen de centrale opslag?	Je noemt hier een van de mogelijke kenmerken van een werksysteem. Hebben we het hier over synchronisatie tussen twee fysieke locaties of een one to many synchronisatie (1:n)	Ik gebruik deze definitie voor synchroniseren: Data naar een tweede (of meer dan dat) locatie repliceren. Dat kan een eenmalige actie zijn (een kopie), het kan realtime zijn, waarbij de beide omgevingen voortdurend gelijk zijn (spiegelen) of dat data op gezette tijden wordt gerepliceerd. Voor bijvoorbeeld HPC is het noodzakelijk om ook terug te kunnen repliceren. Scope staat gedefinieerd in het bijgeleverde worddocument.
3	Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.	Niet alleen een (gesynchroniseerde) backup is gewenst, maar ook een aantal historische versies, om verkeerde verwerkingsactie terug te draaien.	Onduidelijk. Als het puur gaat over backup van data (naar 2 of meer andere veiligheidsopslagen) dan is dit op dit niveau teveel implementatie detail. Dan is de dienstverlening puur: zorgen voor veilige opslag, met geen kans op dataverlies. Dit is niet een specifieke dienst, maar meer een eigenschap van de opslag dienst? Maar hier praat je ook over synchroniseren. Tussen wat dan? De decentrale datahouders en het centrale systeem? Dan lijkt het erg op de vorige dienstverlening.	Wellicht bedoel je het anders, is de term backup ongelukkig gekozen, en is de term 'kopie' beter op zijn plaats. Synchroniseren is niet hetzelfde als een backup. Als de bron data verliest wordt dat via sync ook automatisch verwijderd op de centrale opslaglocatie. Bovendien heeft een backup alleen zin als het werksysteem consistent blijft. *'Data backup is the process of making copies of your data, whatever it may be, so that if your original data is damaged or lost, it's not gone forever. It means to literally have a "backup" in case things go wrong. You can then restore your data back to the same or different location'*	In dit geval worden periodieke syncs bedoeld. Wellicht zou periodieke synchronisatie gecombineerd met versiebeheer goed zijn als backup?
4	Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel (sneller dan tape) opgehaald kunnen worden.	In het algemeen is dit zeker de wens. Bij enorme hoeveelheden data, bijvoorbeeld bij klimaatonderzoek, hebben de meeste onderzoekers begrip voor tape snelheden.	Op dit abstractieniveau zou ik woorden als Tape niet gebruiken. Je mengt hier twee zaken: opslagperiode en snelheid van kunnen gebruiken. Splitsen. En indien mogelijk SMART. En zijn dit echt diensten? Ik denk het niet.	Er vanuitgaand dat op een later tijdstip vastgesteld wordt wat alle variablene in dit statement praktisch betekenen ('relevant' en 'voldoende' zijn rekbare begrippen en zeker niet SMART)	SMARTer dan de lit en interviews mag ik het niet maken. Relevant betekent hier dat de wetenschapper bepaalt wat een geschikte periode is. Bij data bepaalt de wetenschapper welke data relevant zijn om het onderzoek te kunnen herhalen. Dat zal per onderzoek anders kunnen zijn. Tape kan er uit, alleen wordt het dan lastig iets over de snelheid te zeggen.

Tabel 115, Commentaar en reactie over de categorie opslag bij de eerste ronde

Commentaar en reactie over de categorieën relevante data en interactie bij de eerste ronde					
		Commentaar 1 na eerste ronde	**Commentaar 2 na eerste ronde**	**Commentaar 3 na eerste ronde**	**Commentaar en voorstellen Bert Kuipers**
	Relevante data				
5	Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek aan te houden waarbij formats in templates beschreven kunnen worden.	De naamgeving en folderstruuctuur wordt vaak gebruikt om metadata vast te leggen.	Folderstructuur is implementatie. Misschien willen ze in de toekomst wel in een grote logische database met een entiteitenboom werken in de toekomst. Of een data-lake met beperkte structuur. Zou structuur (opslag waarin de samenhang van verschillende datasets, hun onderlinge relaties, en relaties met metadata en data-documentatie) als term definiëren en gebruiken. En dan hier aangeven dat er verplichte, geadviseerde of vrije structuren kunnen worden gebruikt. Ik zie hier trouwens eigenlijk 2 diensten: de mogelijkheid te specificeren. En dan bij gebruik het af te dwingen of te adviseren.	Volgens mij is 'folderstructuur' een nogal ruwe invulling van de F van FAIR (zie ook: https://www.go-fair.org/fair-principles/) en de vraag die rijst is of dat de huidige situatie is, de gewenste of beiden?	Naamconventies en folderstructuren (een hiërarchie van mappen) is iets wat alle wetenschappers in meer of mindere mate gebruiken om hun data vindbaar en volgbaar te houden. Uit de interviews bleek dat dit samen met het ELN het meest hiervoor werd gebruikt. Voor de FAIR opmerking verwijs ik naar het Word document.
6	Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.	Voor software/scripts worden git platforms steeds populairder. Mede hierdoor, groeit het interesse in versie beheer voor datasets.	Voelt een beetje meer als 'opslag'. Misschien de waarde meer specificeren als P&S: de mogelijkheid om data die veranderd is door bewerkingen, meer te kunnen herstellen.	Ik zou zeggen, bestanden, data, programmatuur, (procedures?)	Versiebeheer is specifiek voor volgbaarheid krachtig. Code is onderdeel van onderzoeksdata en valt daardoor onder de relevante data definitie. I.p.v. data en bestanden, herformuleren naar relevante data zou mijn voorstel zijn.
7	Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.	Jazeker, maar verwijzingen door externe (bron)data is ook nodig, bijvoorbeeld in de chemie zijn er internationale databases met moleculaire structuur en eigenschappen. Deze data worden niet opgeslagen bij de eigen onderzoek, maar zijn wel degelijk input data die nodig zijn om een onderzoek te herhalen.	Gezamenlijk' is wat ambigu en neigt naar 'gelijktijdig opslaan', maar dat bedoel je niet. Opslaan van data, waarbij bij het opslaan ook relaties worden geëxpliciteerd met andere al eerder opgeslagen data en meta-data. Dit voelt overigens niet als een dienst.	Zie vorige opmerking	Misschien is in samenhang beter dan gezamenlijk.
8	Het *werksysteem* maakt het mogelijk data te vernietigen en logt deze activiteit.	Ik kan een 'use case' bedenken, maar ik heb persoonlijk deze vraag niet gekregen.	Wellicht zou hier ook iets over (juridisch) bewaartermijnen van toepassing zijn?	Data, kopieën van data alsmede backups (denk aan eventuele Privacy aspecten)	Ook hier weer relevante data. Voorstel: relevante data en alle afgeleiden hiervan.
9	Het *werksysteem* biedt de mogelijkheid om geautomatiseerd *metadata* aan te maken, bij voorkeur m.b.v. een template.	'Geautomatiseerd' betekent dan voor mij 3 dingen: de computer genereerd de metadata zonder inspanning van de onderzoeker; de onderzoeker schrijft scripts die metadata genereerd; de onderzoeker wordt automatisch geconfronteerd met een (maatwerk) formulier om metadata in te vullen.	Ik heb hier geen beeld bij. Bedoel je bijv. een DDL en dan een script die bijv. een tabel aanmaakt met attributen?	… bij voorkeur geautomatiseerd op basis van een template of algoritmen?	Op basis van het antwoord van commentator 1 en aansluitend op nummer 3 wil ik het volgende voorstellen: Het werksysteem genereert de metadata zonder inspanning van de onderzoeker; Het werksysteem voorziet in scripts die metadata genereren; Het werksysteem confronteert een onderzoeker automatisch met een (maatwerk) formulier om metadata in te vullen.
	Interactie				
10	Het *werksysteem* biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).	Een voorbeeld van klimaatonderzoek is dat mensen een userinterface hebben voor de metadata, waarna ze een API gebruiken om de data op te halen.	Interactie is verwarrend. Ik denk dat ik eerder de term 'kanaal' zou gebruiken: Diensten die dit werksysteem aanbiedt aan gebruikers, zijn te gebruiken in een mens-computer interface en een computer-computer interface. Maar… zijn alle diensten alleen IT-geleverd? In feite is dit geen zelfstandig product of service, maar een eis aan/eigenschap van meerdere products and services.		T.a.v. 2: de definitie van products and services is heel breed (zie worddocument). Een interface levert informatie(producten). Ik heb zelf geen beeld bij kanaal. Wellicht herformuleren naar het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via…

Tabel 116, Commentaar en reactie over de categorieën relevante data en interactie bij de eerste ronde

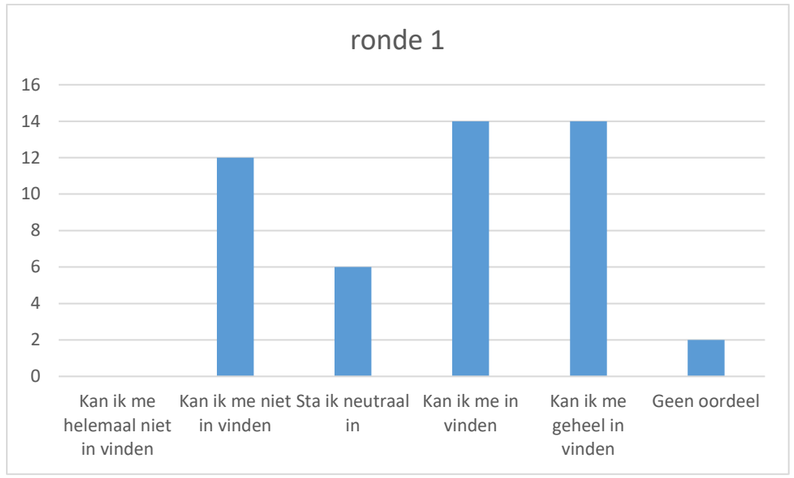
Commentaar en reactie over de categorieën beveiligen en ELN bij de eerste ronde					
		Commentaar 1 na eerste ronde	**Commentaar 2 na eerste ronde**	**Commentaar 3 na eerste ronde**	**Commentaar en voorstellen Bert Kuipers**
	Beveiligen				
11	Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek te beveiligen.	Vooral het woord 'mogelijkheid' vindt ik belangrijk. Soms is het juist de bedoeling om de data open te stellen, bijvoorbeeld voor 'Citizen Science'.	Beveiligen tegen wat? Handig om dit te specificeren. Is dit een losse dienst, of in feite integrale eigenschap van opslag? Of zie je het echt als een extra handeling die een gebruiker kan doen?	Ik zou iets zeggen aangaande de mate van beveiliging, zoals: … in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.	Beveiliging is per onderzoek verschillend. Voor het ene onderzoek gaat het om autorisaties, voor het andere om encryptie. Soms is een NDA achtig contract noodzakelijk. Dit staat wel beschreven in de beveiligingsparagraaf. In lijn met de vertrouwelijkheid van het onderzoek beveiligen lijkt me een goede suggestie.
12	Het *werksysteem* verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.	Hier heb ik nooit een vraag over gekregen, omdat onderzoekers hier vanuit gaan.	Hm. Je hebt diverse (ook ISO!) normen voor kwaliteitseisen rond data. Dit lijkt een selectie hieruit. https://iso25000.com/index.php/en/iso-25000-standards/iso-25012 Ik denk overigens niet dat een worksystem in deze scope dit kan garanderen. Consistentie gaat ook over samenhang van data, en iemand kan een dataset verwijderen waardoor consistentie verloren gaat. Ik zou deze termen scherp definiëren. Idem: voelt niet als individueel product of service.	Ik zou zeggen het geheel accuraat en consistent moet blijven (data, programmatuur, procedures?)	Hier had moeten staan relevante data (inderdaad het geheel). Het samenspel van onderzoeksdata, metadata en datadocumentatie. Wellicht is garanderen een te grote eis. Als ik het verander in bewaakt de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden. Zou dat beter zijn?
	ELN				
13	Het *werksysteem* biedt de mogelijkheid een *ELN* te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.	In het kader van 'reproducibility' van onderzoek, zie ik steeds meer belangstelling hiervoor.	Je mixt twee requirements. 1 lijkt erop dat het werksysteem functionaliteiten aanbiedt voor een ELN (of bedoel je dat het werksysteem kan integreren met een bestaand ELN product??). De andere gaat over dat je die functionaliteiten via verschillende devices kunt gebruiken? Splitsen. Die devices, dat geldt toch ook voor alle andere functionaliteiten? Dan zou ik multidevice bij interactie/kanalen zetten. Is dan ook geen aparte dienst maar een eigenschap.	…. Waarbij de inhoud van het ELN platform-onafhankelijk (verschillende operating systems) bruikbaar is en op verschillende typen devices (smartphones, tablets, laptops, Pc's, smartwatch (?), other) beschikbaar ….'	
14	Het *werksysteem* biedt een *ELN* aan dat de mogelijkheid heeft om o.a. activiteiten en keuzes te loggen.	Dit is de reden onderzoekers een (papier of elektronisch) lab notebook gebruiken. Voor zichzelf, of om het werk van studenten/PhD kandidaten te monitoren.	Een dienst met het woord o.a. is linke boel.	o.a. impliceert dat er meer kan zijn. Zijn dit de minimum eisen aan een ELN?	De suggestie van nummer 2 klinkt logisch. Het werksysteem kan integreren met bestaande ELN vormen. Daar kom dan bij: Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen. De multiplatform eis, is iets wat voor alle IT onderdelen van het ELN geldt. Het is een architectuurprincipe. Ontwerpprincipes (de kruisbestuiving tussen het WSS en de architectuurprincipes) is een volgende stap in het ontwerp. Die zou ik nu nog niet willen toevoegen.
15	Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om te linken naar relevante data in het lopende onderzoek.	Jazeker. Kleine hoeveelheden data moeten ook direct in de ELN opgeslagen kunnen worden. Bijvoorbeeld de kamertemperatuur en barometrische gegevens bij een experiment.	Wat te linken? Onduidelijk. Ik neem aan dat je bedoelt: bij het beheren van data-documentatie en logboek?		Zie hierboven
16	Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden.	Voorbeelden zijn: tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.	Erg globaal als dienst.		Hier wordt met linken bedoeld: verwijzen met een shortcut of verwijzen met een beschreven bestandslocatie (dit staat elders in het onderzoek uitgelegd).

Tabel 117, Commentaar en reactie over de categorieën beveiligen en ELN bij de eerste ronde

15.2.3. Conclusies bij de eerste ronde

Conclusies over scores

Drie experts kunnen maximaal 48 verschillende scores geven als ze per product/dienst een verschillende mening geven. In de eerste ronde werden er 38 van de 48 mogelijkheden gebruikt (zie ook TABEL 114). Hun commentaar wordt beschreven in TABEL 115, TABEL 116 en TABEL 117.
De verdeling van de mening van de experts staat beschreven in FIGUUR 23.



Figuur 23, verdeling scores na eerste ronde

Commentaar bij de eerste ronde

Buiten de commentaren zoals benoemd in TABEL 115, TABEL 116, en TABEL 117 gaf expert bovendien onderstaand commentaar na de eerste ronde:

Ik worstel met je scope afbakening/systeem grenzen. In de grote tabel lijkt het alsof je de wetenschappers onderdeel ziet van het werksysteem (zie activities en partipants), maar in de eisen hierboven gaat het om een werk systeem dat diensten (IT gedreven) aanbiedt aan wetenschappers. Zijn ze dan onderdeel of zijn ze deel van de omgeving die het systeem gebruiken en hier waarde van krijgen? Ik zou gaan voor gebruikers buiten het systeem. Dat betekent dat de in het werksysteem getriggerd worden door een gebruiker en dan zijn de activiteiten die je noemt extern aan het systeem, gedaan door gebruikers, waarbij het systeem bepaalde technische acties doet om dit mogelijk te maken/uit te voeren. Nu staan er in de tabel diverse activiteiten die me meer lijken als extern gedrag dat bepaalde triggers zijn voor activiteit in het werksysteem. De vraag is dan: wat moet het werksysteem doen bij deze triggers? Mijn tip hierbij is: maak een extra plaatje met externe actoren, hun processen en vandaar uit job to be done, de dienst-'ingangen' van het werksysteem en evt. achterliggende processen. Daarmee creëer je een heel scherp beeld van je systeemgrenzen.

Tweede is dat ik heel verschillende interpretaties zie van het woord 'product of dienst'. Soms zie ik pure diensten, maar ik zie ook eigenschappen van diensten. Een handig middel is het 'job to be done raamwerk' : welke behoefte heeft een actor (buiten het werksysteem), welk doel wil deze bereiken, en wat biedt het werksysteem dan precies aan, zodat de persoon dit doel (job to be done) kan bereiken, middels / geholpen door het werk systeem.

Is het mogelijk je diensten te koppelen aan de doelen van het werksysteem. Die mis ik nog. Elk werksysteem heeft een purpose. Iets rond ontzorgen wetenschappers, borgen van data-kwaliteit, …. Dan zie je mogelijk ook diensten die je nog niet hebt gedefinieerd (een doel zonder dienst...)

Conclusie commentaar bij de eerste ronde

Veel van het commentaar heeft met de scope te maken. Die komt tot uitdrukking in de probleemstelling. Ook wordt er regelmatig om meer specificatie gevraagd (SMART). Daar staat tegenover dat de mate van specificatie afhangt van de gevonden begrippen in de literatuur en de uitkomsten van het onderzoek. Er is in de formulering van de products and services getracht, allen op basis daarvan tot uitspraken te komen. Ten slotte blijkt dat de experts in een aantal gevallen een ander idee hebben bij werksysteem en onderdelen van het werksysteem (expert 3 ziet de wetenschappers niet als onderdeel van het werksysteem en expert 2 heeft een ander beeld bij diensten). Voor de tweede ronde wordt een worddocument met extra informatie geleverd en zet ik in een extra kolom commentaar dat vragen zou moeten beantwoorden.

15.2.4. Tweede ronde scores

Scores na de tweede ronde							
		Kan ik me helemaal niet in vinden	**Kan ik me niet in vinden**	**Sta ik neutraal in**	**Kan ik me in vinden**	**Kan ik me geheel in vinden**	**Geen oordeel**
	Opslag						
1	Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.				123		
2	Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.				2	13	
3	Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.			2	13		
4	Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel (sneller dan tape) opgehaald kunnen worden.			2	13		
	Relevante data						
5	Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek aan te houden waarbij formats in templates beschreven kunnen worden.			3	2	1	
6	Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.				3	12	
7	Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.				123		
8	Het *werksysteem* maakt het mogelijk data te vernietigen en logt deze activiteit.				1	23	
9	Het *werksysteem* biedt de mogelijkheid om geautomatiseerd *metadata* aan te maken, bij voorkeur m.b.v. een template.				13		2
	Interactie						
10	Het *werksysteem* biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).				2	13	
	Beveiligen						
11	Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek te beveiligen.			2		13	
12	Het *werksysteem* verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.				2	13	
	ELN						
13	Het *werksysteem* biedt de mogelijkheid een *ELN* te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.				23	1	
14	Het *werksysteem* biedt een *ELN* aan dat de mogelijkheid heeft om o.a. activiteiten en keuzes te loggen.			2	3	1	
15	Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om te linken naar relevante data in het lopende onderzoek.				123		
16	Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden.			2		13	

Tabel 118, Scores na de tweede ronde

15.2.5. Commentaar bij de tweede ronde

Commentaar en reactie over de categorie opslag bij de tweede ronde					
		Commentaar en voorstellen na eerste ronde Bert Kuipers	**Commentaar 1 na tweede ronde**	**Commentaar 2 na tweede ronde**	**Commentaar 3 na tweede ronde**
	Opslag				
1	Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	Voor scope en beschrijving wil ik graag verwijzen naar het bijgeleverde Word document. T.a.v. de 8TB lees ik dat iedereen dat te specifiek vindt. Mijn voorstel zou zijn om dat er uit te halen en te vervangen voor voldoende (door de wetenschapper te bepalen) opslagruimte. T.a.v. van nummer 3: wetenschappers zijn onderdeel van het werksysteem (participants en customers). T.a.v. nummer 2, ik kan niet specifieker worden dan de literatuur en interviews me aan ruimte geven. Ik kan wel anders formuleren.	Eens met je voorstel over de 8TB	Eens met je voorstel. Gekke vraag: ik zie nergens de dienst 'kunnen raadplegen van data'. Mist die dienst of mis ik iets? Idem voor opslaan en muteren.	Voldoende toegelicht
2	Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.	Ik gebruik deze definitie voor synchroniseren: Data naar een tweede (of meer dan dat) locatie repliceren. Dat kan een eenmalige actie zijn (een kopie), het kan realtime zijn, waarbij de beide omgevingen voortdurend gelijk zijn (spiegelen) of dat data op gezette tijden wordt gerepliceerd. Voor bijvoorbeeld HPC is het noodzakelijk om ook terug te kunnen repliceren. Scope staat gedefinieerd in het bijgeleverde worddocument.	Goed om de definitie van 'synchroniseren' te noemen. Ik dacht alleen aan Dropbox-achtige synchroniseren		Voldoende toegelicht
3	Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.	In dit geval worden periodieke syncs bedoeld. Wellicht zou periodieke synchronisatie gecombineerd met versiebeheer goed zijn als backup?	Ja, het noemen van meerdere versies lijkt me belangrijk	Ik zou 'opslaglocatie' (vorige punt) differentieren van 'backup locatie' of voorziening.	Versiebeheer is een vereiste indien je de techniek van synchronisatie gebruikt, alleen op die manier creëer je een integer point-in-time herstelpunt. In de voorgestelde combinatie dus akkoord.
4	Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel (sneller dan tape) opgehaald kunnen worden.	SMARTer dan de lit en interviews mag ik het niet maken. Relevant betekent hier dat de wetenschapper bepaalt wat een geschikte periode is. Bij data bepaalt de wetenschapper welke data relevant zijn om het onderzoek te kunnen herhalen. Dat zal per onderzoek anders kunnen zijn. Tape kan er uit, alleen wordt het dan lastig iets over de snelheid te zeggen.	(sneller dan tape) vervangen door (bepaald door de onderzoeker)?	SMART: gebruiker bepaalt. Snelheid: wellicht direct beschikbaar standaard snelheid vs. indirect beschikbaar (moet worden ingeladen)	Als je voor periode en data de term *relevant* gebruikt dan kun je tape wellicht vervangen door: .. ... de relevante data met voldoende snelheid kan worden opgehaald. Hierbij rijst nog wel de vraag of alle data in het werksysteem met dezelfde snelheid moet kunnen worden opgehaald of dat hier sprake kan zijn van differentiatie.

Tabel 119, Commentaar en reactie over de categorie opslag bij de tweede ronde

Commentaar en reactie over de categorieën relevante data en interactie bij de tweede ronde					
		Commentaar en voorstellen na eerste ronde Bert Kuipers	**Commentaar 1 na tweede ronde**	**Commentaar 2 na tweede ronde**	**Commentaar 3 na tweede ronde**
	Relevante data				
5	Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek aan te houden waarbij formats in templates beschreven kunnen worden.	Naamconventies en folderstructuren (een hiërarchie van mappen) is iets wat alle wetenschappers in meer of mindere mate gebruiken om hun data vindbaar en volgbaar te houden. Uit de interviews bleek dat dit samen met het ELN het meest hiervoor werd gebruikt. Voor de FAIR opmerking verwijs ik naar het Word document.	-	Als dit taal en concepten wetenschappers zijn, dan lijkt me dit voldoende verwoord. Ik mis je reactie op splitsing van diensten, maar kan leven met de combinatie.	Blijft mijn vraag of je hier de huidige of de gewenste situatie beschrijft. Wat is je uitgangspunten hierbij?
6	Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.	Versiebeheer is specifiek voor volgbaarheid krachtig. Code is onderdeel van onderzoeksdata en valt daardoor onder de relevante data definitie. I.p.v. data en bestanden, herformuleren naar relevante data zou mijn voorstel zijn.	Eens met je herformulering	Eens met data als generieke term (wel ergens anders dan definiëren als aggregaat). Ik interpreteer je requirement nu anders: het gaat om versies na een specifieke bewerking (zodat je terug kunt gaan).	Akkoord met de voorgestelde aanpassing.
7	Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.	Misschien is in samenhang beter dan gezamenlijk.	Eens met je herformulering	Tekstvoorstel: In samenhang met interne en externe bronnen, inclusief metagegevens die deze samenhang beschrijven	Akkoord met de voorgestelde aanpassing.
8	Het *werksysteem* maakt het mogelijk data te vernietigen en logt deze activiteit.	Ook hier weer relevante data. Voorstel: relevante data en alle afgeleiden hiervan.	Mooie herformulering	Eens , inclusief afgeleiden ervan (en dus ook backups)	Akkoord met de voorgestelde aanpassing.
9	Het *werksysteem* biedt de mogelijkheid om geautomatiseerd *metadata* aan te maken, bij voorkeur m.b.v. een template.	Op basis van het antwoord van commentator 1 en aansluitend op nummer 3 wil ik het volgende voorstellen: Het werksysteem genereert de metadata zonder inspanning van de onderzoeker; Het werksysteem voorziet in scripts die metadata genereren; Het werksysteem confronteert een onderzoeker automatisch met een (maatwerk) formulier om metadata in te vullen.	Misschien 'en/of' steeds toevoegen? 1 of 2 of 3 van de opties kan voldoende/nodig zijn	Nog steeds geen beeld bij automatische generatie. Als ik een dataset heb (bijv. tabel met kolommen) dan moet ik toch zelf aangeven wat elke kolom is, en wat de entiteit/tabel bevat?	'scripts' is invulling, ik zou zeggen: 'Het werksysteem levert voorzieningen om metadata te genereren'. Verder zou ik i.p.v. de term confronteert de term faciliteert gebruiken (confronteren heeft voor mij een negatieve associatie
	Interactie				
10	Het *werksysteem* biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).	T.a.v. 2: de definitie van products and services is heel breed (zie worddocument). Een interface levert informatie(producten). Ik heb zelf geen beeld bij kanaal. Wellicht herformuleren naar het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via…	Je herformulering vind ik mooier	Kanaal = interface. Een interface stelt gebruiker in staat een bepaalde dienst af te nemen. Eens met je voorstel: het werksysteem biedt meerdere interfaces om producten & diensten te gebruiken (zoals mens-machine, machine-machine).	Ik heb persoonlijk geen moeite met de oorspronkelijke formulering.

Tabel 120, Commentaar en reactie over de categorieën relevante data en interactie bij de tweede ronde

Commentaar en reactie over de categorieën beveiligen en ELN bij de tweede ronde					
		Commentaar en voorstellen na eerste ronde Bert Kuipers	**Commentaar 1 na tweede ronde**	**Commentaar 2 na tweede ronde**	**Commentaar 3 na tweede ronde**
	Beveiligen				
11	Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek te beveiligen.	Beveiliging is per onderzoek verschillend. Voor het ene onderzoek gaat het om autorisaties, voor het andere om encryptie. Soms is een NDA achtig contract noodzakelijk. Dit staat wel beschreven in de beveiligingsparagraaf. In lijn met de vertrouwelijkheid van het onderzoek beveiligen lijkt me een goede suggestie.	'in lijn met' vind ik een goede toevoeging	Vraag: zou het kunnen dat verschillende data-onderdelen andere beveiligingseisen nodig hebben? Bijv. persoonsgegevens (afschermen!) vs. anonieme afgeleide gegevens(breder te delen) hiervan. Zo ja, dan zou ik dit hier wel benoemen (dus meer granulariteit in beveiliging, per dataset). Zou afschermen en gericht toegang geven aan verschillende doelgroepen als term gebruiken, i.c.m. vertrouwelijkheid van de datasets.	Geen verdere aanvullingen.
12	Het *werksysteem* verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.	Hier had moeten staan relevante data (inderdaad het geheel). Het samenspel van onderzoeksdata, metadata en datadocumentatie. Wellicht is garanderen een te grote eis. Als ik het verander in bewaakt de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden. Zou dat beter zijn?	'relevante data' en 'bewaakt' vind ik beter	Eens met je voorstel, bewaken en signaleren handelingen die inconsistenties gaan veroorzaken (en dit loggen).	Voorgestelde aanpassing is voor mij voldoende.
	ELN				
13	Het *werksysteem* biedt de mogelijkheid een *ELN* te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.	De suggestie van nummer 2 klinkt logisch. Het werksysteem kan integreren met bestaande ELN vormen. Daar kom dan bij: Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen. De multiplatform eis, is iets wat voor alle IT onderdelen van het ELN geldt. Het is een architectuurprincipe. Ontwerpprincipes (de kruisbestuiving tussen het WSS en de architectuurprincipes) is een volgende stap in het ontwerp. Die zou ik nu nog niet willen toevoegen	prima aanpassing	Eens met je voorstel - splitsen. Wat bedoel je met ELN 'vormen'? Toepassingen? Maar integreren is ook een stevige requirement zo - alle ELN-toepassingen? Is er een soort generieke interface voor ELN-toepassingen? Zo nee, dan lijkt me integratie out of the box met 'alle' bestaande ELN's te grote requirement. Misschien wel een top-3 meest gebruikte??	Akkoord met de voorgestelde aanpassing.
14	Het *werksysteem* biedt een *ELN* aan dat de mogelijkheid heeft om o.a. activiteiten en keuzes te loggen.	Zie hierboven	zie hierboven	"En wat verder relevant is": Dit voelt als een gebied dat idd nog niet voldoende is gescoped.	De voorgestelde aanpassing hierboven maakt dat deze regel overbodig lijkt te zijn geworden.
15	Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om te linken naar relevante data in het lopende onderzoek.	Hier wordt met linken bedoeld: verwijzen met een shortcut of verwijzen met een beschreven bestandslocatie (dit staat elders in het onderzoek uitgelegd).	ja, beter om het zo te beschrijven i.p.v. het woord 'linken' te gebruiken	Ok, nu begrijp ik deze beter (en neem ik aan dat zo'n verwijzing dan onderdeel is van ELN-data zoals een activiteit-beschrijving (doe dit met bestand XYZ), keuzes (we hebben bestand ABC(link) verwerkt met code ZZZ(link), …)	Alternatieve omschrijving: *…biedt de mogelijkheid om een link vast te leggen naar relevante data ….*
16	Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden.	Commentator 1 legt goed uit wat hier bedoeld wordt.	-	Gebied dat nog onvoldoende gescoped lijkt. Mijn aanname: dat je als onderdeel van je werksysteem een ELN aanbiedt dat een (digitaal?) platform bevat waarbinnen je een soort 'apps' kan ontwikkelen of integreren gericht op specifieke onderzoekstoepassingen?	Suggestie om voorbeeld toe te voegen, dus: …. voor bepaalde vakgebieden zoals: tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.

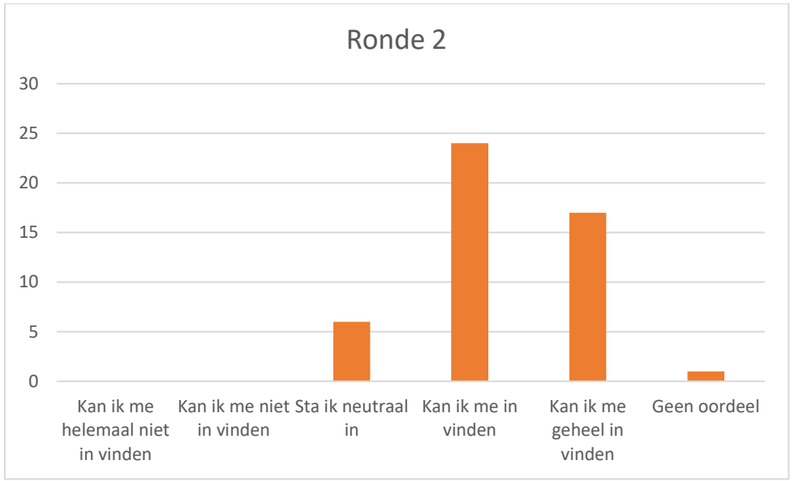
Tabel 121, Commentaar en reactie over de categorieën beveiligen en ELN bij de tweede ronde

15.2.6. Conclusies bij de tweede ronde

Conclusies over scores bij de tweede ronde

Drie experts kunnen maximaal 48 verschillende scores geven als ze per product/dienst een verschillende mening geven. In de tweede ronde werden er 31 van de 48 mogelijkheden gebruikt (zie ook TABEL 118). Dat zijn er zeven minder dan in ronde één. De experts kruipen naar elkaar toe. Bovendien wordt de optie 'kan ik me niet in vinden' niet meer gebruikt. Effectief is er consensus na de tweede ronde. Commentaar bij deze ronde staat beschreven in TABEL 119, TABEL 120 en TABEL 121.

De verdeling van de mening van de experts staat beschreven in FIGUUR 24.



Figuur 24, verdeling scores na tweede ronde

Conclusie commentaar bij de eerste ronde

Commentaar in de tweede ronde heeft vooral te maken met suggestie voor verbeteringen en nog wat extra vragen. Hierop wordt ingegaan in de kolom commentaar en voorstellen in de tabellen: TABEL 119, TABEL 120, en TABEL 121.

15.2.7. Derde ronde scores

Scores na de derde ronde							
		Kan ik me helemaal niet in vinden	**Kan ik me niet in vinden**	**Sta ik neutraal in**	**Kan ik me in vinden**	**Kan ik me geheel in vinden**	**Geen oordeel**
	Opslag						
1	Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, door de onderzoeker te bepalen, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.				13	2	
2	Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.				2	13	
3	Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie, waarbij versies worden bijgehouden van de verschillende replica's en waarbij een bepaalde versie van een replica teruggesynchroniseerd kan worden.				13	2	
4	Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel opgehaald kunnen worden. Voldoende snel is afhankelijk van hoeveelheden relevante data en de mate waarin ze nog gebruikt worden.			2	13		
	Relevante data						
5	Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek te ondersteunen waarbij formats in templates beschreven kunnen worden.			3	2	1	
6	Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op relevante data te kunnen toepassen.				3	12	
7	Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de relevante data zo op te slaan, dat duidelijk is dat ze bij elkaar horen.				13	2	
8	Het *werksysteem* maakt het mogelijk relevante data en alle afgeleiden hiervan, te vernietigen en logt deze activiteit.				1	23	
9	Het werksysteem levert voorzieningen om metadata te genereren met zo min mogelijk inspanning van de onderzoeker, bij voorkeur geautomatiseerd op basis van een template of algoritmen, maar ook met voldoende ruimte voor eigen invulling.				123		
	Interactie						
10	Het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via meerdere interfaces (zoals mens-machine, machine-machine).				2	13	
	Beveiligen						
11	Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.			2		13	
12	Het *werksysteem* bewaakt de integriteit van de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden in logging.			2		13	
	ELN						
13	Het werksysteem kan integreren met bestaande, veelgebruikte, ELN toepassingen.				3	12	
14	Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen.				23	1	
15	Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om een link (verwijzing, 'shortcut') vast te leggen naar relevante data in het lopende onderzoek.				123		
16	Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden, zoals het tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.			2		13	

Tabel 122, Scores na de derde ronde

15.2.8. Commentaar in de derde ronde

In deze ronde geeft alleen expert 2 nog commentaar.

Commentaar en reactie over de categorie opslag bij de derde ronde			
	Uiteindelijke voorstel	**Commentaar Bert Kuipers bij reacties op tweede ronde**	**Commentaar expert 2**
	Opslag		
1	Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, door de onderzoeker te bepalen, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.	Terechte vraag over het raadplegen van de data. Het RDM werksysteem houdt zich bezig met het beheer van de data. De onderzoeker haalt data op, of schrijft het weg of bewerkt het (ophalen en wegschrijven hebben met vindbaarheid te maken, wijzigen met volgbaarheid. Deze handelingen staan bij activiteiten. Deze regel gaat over het opslaan van data en regel 6 over het ophalen van data, regel 12 gaat over vernietigen (en kunnen nazoeken dat dat heeft plaatsgevonden). Hiermee levert de dienst products and services die corresponderen met de activities.	
2	Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.		
3	Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie, waarbij versies worden bijgehouden van de verschillende replica's en waarbij een bepaalde versie van een replica teruggesynchroniseerd kan worden.		

Tabel 123, Commentaar en reactie over de categorie opslag bij de derde ronde

Commentaar en reactie over de categorieën relevante data en interactie bij de derde ronde			
	Uiteindelijke voorstel	**Commentaar Bert Kuipers bij reacties op tweede ronde**	**Commentaar expert 2**
4	Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel opgehaald kunnen worden. Voldoende snel is afhankelijk van hoeveelheden relevante data en de mate waarin ze nog gebruikt worden.	Voor het latere technische ontwerp, suggereert dit verschillende oplossingen met verschillende snelheden.	Suggestie: en afhankelijk van beschikbare technology.
	Relevante data		
5	Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek te ondersteunen waarbij formats in templates beschreven kunnen worden.	Dit is zowel de huidige als de gewenste situatie. Verschillende afdelingen passen dit op verschillende wijze toe. Het werksysteem moet hier voldoende ruimte voor bieden.	Waarbij formats en folderstructuur in templates…
6	Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op relevante data te kunnen toepassen.		
7	Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de relevante data zo op te slaan, dat duidelijk is dat ze bij elkaar horen.	Let op, dit is iets anders geformuleerd dan het voorstel van de tweede ronde. In samenhang met interne en externe bronnen, inclusief metagegevens die deze samenhang beschrijven. Relevante data (datasets, metadata en datadocumentatie) horen bij elkaar in een onderzoek, maar dat is niet per definitie evident. Datadocumentatie kan beschrijven hoe een dataset tot stand is gekomen, metadata kan wat zeggen over de betekenis van de dataset. Wel staat verderop de mogelijkheid om naar data te verwijzen in de datadocumentatie. Als je verwijst naar je dataset en je metadata in je datadocumentatie, heb je samenhang. Een andere manier om de samenhang te bewaren zit in naamgeving en folderstructuur.	
8	Het *werksysteem* maakt het mogelijk relevante data en alle afgeleiden hiervan, te vernietigen en logt deze activiteit.		
9	Het werksysteem levert voorzieningen om metadata te genereren met zo min mogelijk inspanning van de onderzoeker, bij voorkeur geautomatiseerd op basis van een template of algoritmen, maar ook met voldoende ruimte voor eigen invulling.	De suggestie opvolgend in de reactie van commentator 1, de drie voorstellen gecombineerd. Tekst voorstellen van 3 meegenomen. In reactie op nr. 2: Bepaalde meetsystemen hebben software die meteen een metadatatabel kunnen aanmaken (wanneer gemaakt, door wie, de betekenis van de verschillende parameters enz.) Er zijn zelfs systemen die je specifiek voor alleen de generatie van metadata in kunt zetten.	
	Interactie		
10	Het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via meerdere interfaces (zoals mens-machine, machine-machine).		

Tabel 124, Commentaar en reactie over de categorie opslag bij de derde ronde

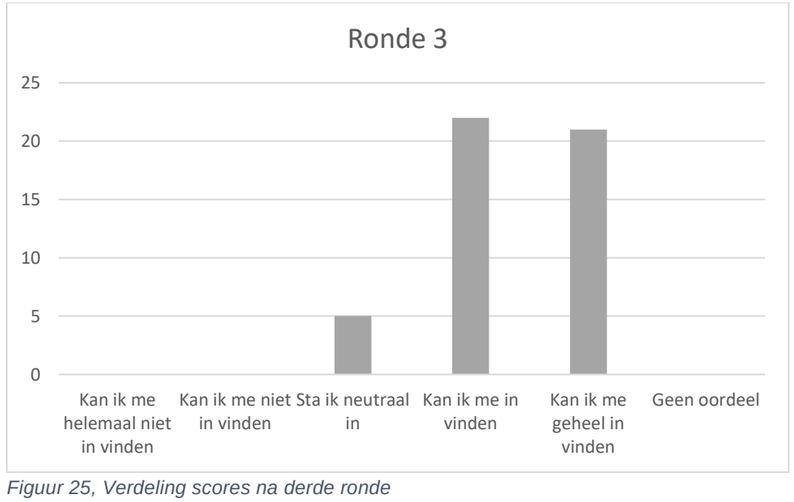
Commentaar en reactie over de categorieën beveiligen en ELN bij de derde ronde			
	Uiteindelijke voorstel	**Commentaar Bert Kuipers bij reacties op tweede ronde**	**Commentaar expert 2**
	Beveiligen		
11	Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.		mis nog steeds granulariteit: bepaalde deelsets die je extra wil afschermen, andere minder. 'Relevante data' denkt dat voor mij nog niet heel scherp.
12	Het *werksysteem* bewaakt de integriteit van de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden in logging.		dan bewaakt het systeem niet, maar logt alleen
	ELN		
13	Het werksysteem kan integreren met bestaande, veelgebruikte, ELN toepassingen.		
14	Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen.		
15	Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om een link (verwijzing, 'shortcut') vast te leggen naar relevante data in het lopende onderzoek.		
16	Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden, zoals het tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.	Het gaat niet zozeer om apps die je ontwikkelt. Het gaat om bepaalde veel voorkomende processen en procedures in een vakgebied die 'voorgekookt' zijn in het ELN. Zo hoeft de onderzoeker dat niet steeds zelf te doen.	

Tabel 125, Commentaar en reactie over de categorieën beveiligen en ELN bij de derde ronde

15.2.9. Conclusies bij de derde ronde

Conclusies over scores bij de derde ronde

Drie experts kunnen maximaal 48 verschillende scores geven als ze per product/dienst een verschillende mening geven. In de derde ronde werden er 31 van de 48 mogelijkheden gebruikt (zie ook TABEL 122). Dat is gelijk aan ronde twee. De experts zijn het ogenschijnlijk iets meer eens geworden, doordat er niemand meer is die geen oordeel geeft en ook iets minder voor de neutrale optie wordt gekozen. Men is het dus wat meer eens met de geformuleerde products and services dan in de tweede ronde (wat logisch is, omdat de suggesties zijn overgenomen). Er is nog steeds consensus, die is zelfs iets sterker geworden. Het commentaar bij deze ronde staat in TABEL 123, TABEL 124 en TABEL 125. De verdeling van de mening van de experts staat beschreven in FIGUUR 25.



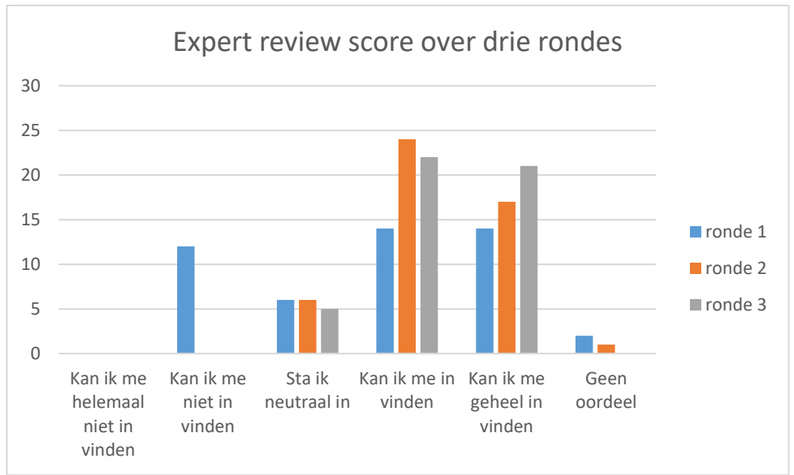
Figuur 25, Verdeling scores na derde ronde

Conclusie commentaar bij de derde ronde

Alleen expert twee geeft nog enkele mogelijke verfijningen aan en heeft enige twijfel bij product/dienst 11. Het zijn op het oog kleine verbeteringen en expert 2 geeft geen negatieve scores. Daarmee wordt de expert review als afgerond beschouwd en rondgestuurd naar de reviewers. Het commentaar van expert 2 staat in de tabellen: TABEL 123, TABEL 124, en TABEL 125.

15.2.10. Conclusie Expert review

De expert review bereikte na de eerste ronde reeds consensus. De tweede ronde werden verfijningsvoorstellen gedaan, die in de derde voor het grootste deel werden bevestigd. Geen van de experts was het oneens met de formulering van de products and services. De review was daarmee tot een einde gekomen. Het verschuiven van de meningen en het toegroeien naar consensus wordt beschreven in FIGUUR 26.



Figuur 26, verschuiven van meningen en toegroeien naar consensus in expert review over drie rondes

De uiteindelijke formulering van de products and services is daarmee:

- Het werksysteem verzekert dat er in het lopende onderzoek voldoende, door de onderzoeker te bepalen, betrouwbare opslagruimte per onderzoek beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.
- Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen onderzoeksdatahouders en een centrale opslag.
- Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie, waarbij versies worden bijgehouden van de verschillende replica's en waarbij een bepaalde versie van een replica teruggesynchroniseerd kan worden.
- Het werksysteem biedt een opslagplaats waar voor een relevante periode de relevante data van een onderzoek voldoende snel opgehaald kunnen worden. Voldoende snel is afhankelijk van hoeveelheden relevante data en de mate waarin ze nog gebruikt worden.
- Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek te ondersteunen waarbij formats in templates beschreven kunnen worden.
- Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op relevante data te kunnen toepassen.
- Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de relevante data zo op te slaan, dat duidelijk is dat ze bij elkaar horen.
- Het werksysteem maakt het mogelijk relevante data en alle afgeleiden hiervan, te vernietigen en logt deze activiteit.
- Het werksysteem levert voorzieningen om metadata te genereren met zo min mogelijk inspanning van de onderzoeker, bij voorkeur geautomatiseerd op basis van een template of algoritmen, maar ook met voldoende ruimte voor eigen invulling.
- Het werksysteem biedt de mogelijkheid om informatie uit te wisselen en opdrachten te ontvangen via meerdere interfaces (zoals mens-machine, machine-machine).
- Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek in lijn met de vertrouwelijkheid van het onderzoek te beveiligen.
- Het werksysteem bewaakt de integriteit van de relevante data op accuraatheid en consistentie door alle bewerkingen bij te houden in logging.
- Het werksysteem kan integreren met bestaande, veelgebruikte, ELN toepassingen.
- Het werksysteem biedt een generiek ELN aan dat het mogelijk maakt activiteiten, ideeën, keuzes en wat verder relevant is voor het onderzoek te loggen.
- Het werksysteem biedt een ELN dat de mogelijkheid heeft om een link (verwijzing, 'shortcut') vast te leggen naar relevante data in het lopende onderzoek.
- Het werksysteem biedt een ELN aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden, zoals het tekenen van molecuulstructuren in de scheikunde en monsterbeheer in de biologie.

15.3. Bijlage correspondentie en bijlagen bij de expert review

Hieronder volgen de mails en de bijlagen die naar de experts zijn gestuurd in de expert review. In de mails wordt aan Excelsheets gerefereerd. Die staan benoemd in de tabellen: TABEL 115, TABEL 116, TABEL 117, TABEL 119, TABEL 120, TABEL 121, TABEL 123, TABEL 124 en TABEL 125.

15.3.1. Mail eerste ronde

Beste Expert,

Hierbij het verzoek tot invulling van de expert review van bevindingen uit mijn afstudeeronderzoek. Ik wil het volgende proces met jou en twee andere (voor jou onbekende) reviewers doorlopen:

- Je krijgt een Word document met korte uitleg over de inhoud van de studie en hetgeen beoordeeld moet worden (bij vragen over de inhoud kun je me benaderen, ik deel dat ook met de anderen);
- Er bestaan in principe geen beperkingen bij de beoordeling, het mag over de inhoud en/of de formulering gaan, wel verzoek ik je vanuit je eigen vakgebied te beoordelen en ook een evidence based beoordeling te geven (zeker als je de voorgestelde regel wilt aanpassen). Dit kan zijn gebaseerd op praktijkervaring en/of onderzoek.
- Je krijgt ook een Excelsheet waarin de beoordeling gegeven kan worden. De beoordeling bestaat uit twee delen:
    - Een score: In hoeverre je je vanuit je expertise kunt vinden in de voorgelegde regel;
    - (evidence based) Uitleg, opmerkingen en/of onderbouwing in tekst.
- Je stuurt de ingevulde Excel naar me terug. Mogelijk vraag ik om wat verduidelijking;
- De scores en de opmerkingen worden met alle reviewers gedeeld. Er zal gevraagd worden (daar waar jullie het oneens zijn) of je je kunt vinden in de uitleg van 1 van de anderen en of je je eerdere mening daarop zou willen bijstellen (eventueel na herformulering van de regel);
- Indien nodig worden regels nu op nieuw geformuleerd en naar de reviewgroep gestuurd, met de bedoeling tot consensus te komen;
- Mocht er voor een bepaalde regel geen consensus worden bereikt, dan wordt deze met uitleg buiten het WSS gehouden.

Ik zou graag de eerste ronde op 8 feb willen afronden, dus indien mogelijk wil ik je antwoorden graag de 5de ontvangen.
De tweede ronde wil ik graag een week later afronden, dus je antwoorden bij voorkeur op de 12de en uiterlijk op de 15de.
De derde ronde volgt weer een week later (bij voorkeur de 19de, maar uiterlijk de 22ste).
Uiteraard zal ik de stukken eerder doorsturen als ik alles binnen heb.

Alvast bedankt voor je medewerking!

Groeten,

Bert

15.3.2. Mail bijlage met extra uitleg bij mail eerste ronde

Expert review RDM werksysteem

Hieronder volgt de probleemstelling van het onderzoek, gevolgd door de meest belangrijke begrippen. In een tekening wordt de samenhang uitgelegd, waarna eerst het volledige WSS (zie hieronder voor uitleg) wordt getoond. Daarna volgen de te beoordelen stellingen met wat extra uitleg indien nodig.

Probleemstelling
Welk ontwerp, omschreven in een worksystem snapshot, beschrijft een Research Data Management werksysteem, toegespitst op het beheer van relevante data voor lopende onderzoeken binnen de case organisatie, zodat deze data vindbaar, volgbaar en voor relevante periode op te slaan zijn?

Belangrijke begrippen voor de probleemstelling en het Werksysteem

Work system (werksysteem): "Een werksysteem is een systeem waarin menselijke participants en/of machines werk (processen en activiteiten) uitvoeren met behulp van informatie, technology en andere middelen om specifieke products and services te produceren voor specifieke interne en/of externe customers" Voor meer informatie kun je bijvoorbeeld kijken op: https://repository.usfca.edu/cgi/viewcontent.cgi?article=1034&context=at

Work System Snapshot: "Het WSS is een samenvatting van één pagina van een werksysteem dat de belangrijkste componenten van zes centrale elementen van het werksysteem identificeert".

Research Data Management (RDM): 'Datamanagement is kort samengevat het creëren, opslaan, onderhouden, beschikbaar maken, archiveren en langdurig bewaren van onderzoeksdata. Hierbij wordt als einddoel vaak gerefereerd aan de zogenaamde FAIR principes: Findable, Accessible, Interoperable and Reusable'.

Relevante data: Alle data (onderzoeksdata, metadata en datadocumentatie) kunnen relevant zijn voor het herhalen van het onderzoek (of van een deel van het onderzoek) door dezelfde of door een andere onderzoeker. Welke data daarvoor precies relevant zijn, is ter beoordeling van de oorspronkelijke onderzoeker(s).

> Onderzoeksdata: 'feitelijke data (zoals numerieke scores, tekstuele records, afbeeldingen en geluiden) die worden gebruikt als primaire bronnen voor wetenschappelijk onderzoek en die algemeen aanvaard zijn in de wetenschappelijke gemeenschap om de onderzoeksresultaten te valideren'.
>
> Een belangrijke constatering uit de interviews is dat onderzoeksdata op drie niveaus bestaan:
>
> 1. Primair: De data zoals die direct na ontstaan zijn verkregen;
> 2. Intermediate: De data die bewerkt zijn en waar analyses op verricht worden;
> 3. Final: De data die gebruikt worden om de tabellen en figuren in de publicatie te vullen.
>
> Metadata zijn data die informatie geven over de onderzoeksdata met als doel de onderzoeksdata voor anderen bruikbaar te maken (reproduceerbaar en interpreteer baar). Bovendien kunnen metadata een indicatie geven van de waarde van de verzamelde gegevens en daarmee mogelijk een juiste herhaling van het onderzoek bevorderen.
>
> Het lijkt belangrijk voor de herhaalbaarheid van het onderzoek dat, buiten de informatie die de metadata al bieden, bekend is hoe de data op niveau 1 zijn verzameld en hoe ze zijn bewerkt om op niveau 2 en 3 te komen. Dit wordt inhoudelijk beschreven in datadocumentatie:
>
> Datadocumentatie tijdens onderzoek betekent het georganiseerd bijhouden van aantekeningen over hoe de data zijn verzameld, wat de resulterende databestanden zijn en hoe ze zijn verwerkt. (DATADOCUMENTATIE CONCLUSIE).

Vindbaarheid (voor dit onderzoek): De stakeholders in de managing active data fase van een onderzoek weten waar de data van een onderzoek staat. De bijbehorende vraag is: "Waar zijn de data (fysiek en/of logisch)?"

Volgbaarheid omvat de data provenance, het proces van het bijhouden van wijzigingen in de onderzoeksdata, inclusief de middelen waarin die wijzigingen worden bijgehouden. De bijbehorende vraag is: "Kan ik de wijzigingen die de data ondergaan volgen op basis van de beschikbare informatie?"

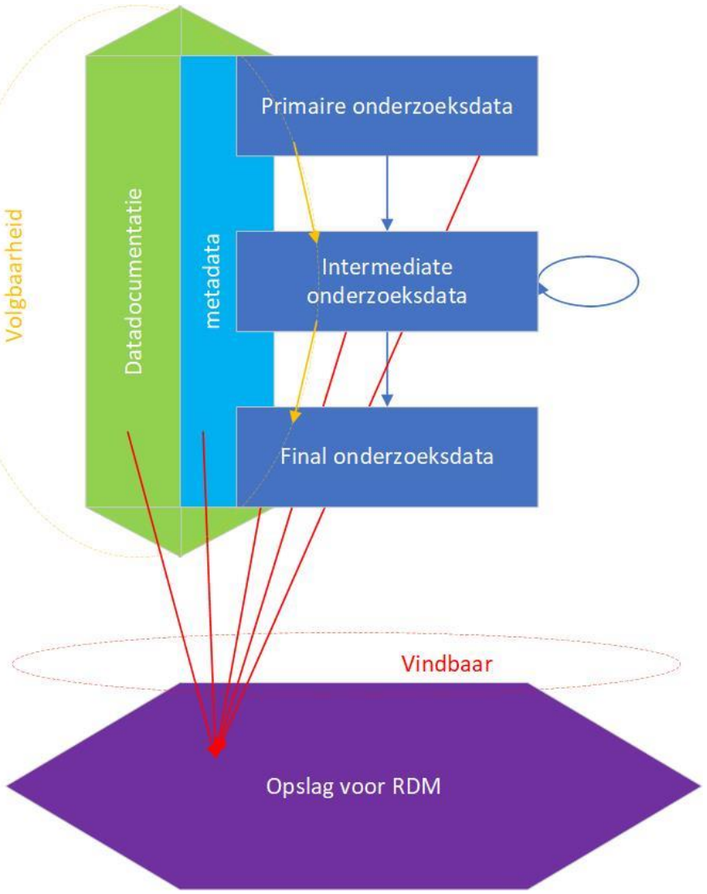
Synchroniseren: Data naar een tweede (of meer dan dat) locatie repliceren. Dat kan een eenmalige actie zijn (een kopie), het kan realtime zijn, waarbij de beide omgevingen voortdurend gelijk zijn (spiegelen) of dat data op gezette tijden wordt gerepliceerd.

Relevante periode: Veel data worden opgeslagen op externe datahouders (denk aan meetsystemen, HPC systemen, werkstations e.d.) en meer gecentraliseerde opslaglocaties (o.a. projectdrive, DropBox, SurfDrive, Home en group drives). Het wordt pas weggegooid als het niet meer nodig is, te bepalen door de wetenschapper. Analoog aan relevante data spreken we van een relevante periode voor de opslagduur.

Onderzoeksdatahouders: Alle decentrale devices en of systemen (en mogelijk hun backups naar lokale media) waar onderzoeksdata op verzameld worden (ELN's, HPC omgevingen, meetinstrumenten, laptops, desktops, flashdrives enzovoort).

ELN: Electronic Lab Notebook, de digitale vorm van het papieren lab logboek. De meeste p&s die over het ELN gaan, gelden zowel voor de electronische als de papieren versie.

In één tekening samengevat ziet dat er als volgt uit:


Volgbaarheid
Datadocumentatie
metadata
Primaire onderzoeksdata
Intermediate onderzoeksdata
Final onderzoeksdata
Vindbaar
Opslag voor RDM

<table>
<tr><th colspan="3">Work system snapshot voor managing active data die vindbaar, volgbaar en archiveerbaar zijn</th></tr>
<tr><th>Customers</th><th colspan="2">Products & Services</th></tr>
<tr><td>• Principle Investigators (PI’s);
• Onderzoekers;
• Computational stakeholders.</td><td colspan="2">• Het werksysteem verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de relevante data benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.
• Het werksysteem levert functionaliteit om data te synchroniseren (in beide richtingen) tussen externe datahouders en een centrale opslag.
• Het werksysteem biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de relevante data in het lopende onderzoek aan te houden waarbij formats in templates worden beschreven.
• Het werksysteem biedt de mogelijkheid een ELN te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.
• Het werksysteem verzekert dat relevante data van het lopende onderzoek vindbaar en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.
• Het werksysteem verzorgt geautomatiseerd een backup van de files en folders door te synchroniseren (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.
• Het werksysteem biedt de mogelijkheid om geautomatiseerd metadata aan te maken, bij voorkeur m.b.v. een template.
• Het werksysteem biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).
• Het werksysteem maakt het mogelijk data te vernietigen en logt deze activiteit.
• Het ELN in het werksysteem biedt de mogelijkheid om o.a. activiteiten en keuzes te loggen
• Het ELN in het werksysteem biedt de mogelijkheid om te linken naar relevante data in het lopende onderzoek.
• Het werksysteem levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.
• Het werksysteem verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.
• Het werksysteem biedt de mogelijkheid de relevante data in het lopende onderzoek te beveiligen.
• Het ELN in het werksysteem kan specifiek gemaakt worden voor bepaalde vakgebieden
• Het werksysteem biedt een opslagplaats voor langdurige relevante data die voldoende snel (sneller dan tape) opgehaald kan worden</td></tr>
<tr><th colspan="3">Major activities and processes</th></tr>
<tr><td colspan="3">• Een researcher vraagt opslagruimte voor zijn onderzoeks(meta)data aan
• Het werksysteem keurt de aanvraag en kent ruimte toe of niet;
• Idem aan de vorige twee bullets, maar voor een tussentijdse aanvraag voor extra ruimte.
• De researcher verplaatst (meta)data op de geboden opslagruimte;
• De onderzoeker ruimt data op
• De onderzoeker verwijst naar een resulterend databestand met een link in de datadocumentatie
• De onderzoeker slaat voldoende relevante data uit het lopende onderzoek op op een plek waar hij/zij het kan terugvinden
• De onderzoeker bewerkt onderzoeksdata op 3 verschillende niveaus
• De onderzoeker maakt metadata aan/gebruikt metadata
• De onderzoeker genereert geautomatiseerd metadata
• De onderzoeker beschrijft metadata in lablogboeken
• De PI adviseert een folderstructuur of legt deze op
• De PI adviseert een naamconventie of legt deze op
• De onderzoeker gebruikt een folderstructuur
• De onderzoeker gebruikt een naamconventie
• De onderzoeker houdt georganiseerd aantekeningen bij over de resulterende databestanden (wat ze zijn, waar ze staan e.d.)
• Er worden backups gemaakt van relevante data voor het lopende onderzoek
• De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verzameld
• De onderzoeker houdt georganiseerd aantekeningen bij over hoe de onderzoeksdata zijn verwerkt
• De relevante data bewerkingen en benaderingen worden gelogd
• De onderzoeker pas versiebeheer toe
• De onderzoeker herhaalt (delen van) het onderzoek</td></tr>
<tr><th>participants</th><th>Information</th><th>Technologies</th></tr>
<tr><td>• Principle Investigators (PI’s);
• Onderzoekers;
• Data Stewards;
• IT ondersteuners (in de afdeling/sectie).
• Computational stakeholders.</td><td>• Adviezen van data stewards aan onderzoekers (inclusief het opmaken van een data management plan);
• Het data management plan, met de planning van het databeheer.
• Adviezen van senior researchers (meestal de PI) aan de onderzoekers (o.a. voor folderstructuren en naamconventies);
• Versiebeheer handmatig uitgevoerd;
• Ruimte aanvragen door een PI;
• Overzichten uit logging.</td><td>• De diverse toegepaste opslagsystemen (zoals benoemd in Data opslag conclusie);
• Data delen vanaf de verschillende toegepaste opslagsystemen (zoals benoemd in Data delen conclusie);
• Data beveiligingstechnieken (zoals benoemd in Conclusie data beveiliging);
• Versiebeheer geautomatiseerd uitgevoerd;
• ELN’s (en ook papieren logboeken).</td></tr>
</table>

Hieronder staan de te beoordelen products and services (p&S) herhaald. De volgorde is alleen veranderd om ze in logische categorieën te plaatsen (in het WSS stonden ze op volgorde van vindbaarheid, volgbaarheid en opslag voor een relevante periode). Begrippen die aan het begin van dit document staan beschreven, zijn *cursief* gemaakt. Indien nodig staat er wat uitleg bij de regel.

De meeste p&s zijn het gevolg van literatuuronderzoek en verdere specificaties na interviews bij de case organisatie. De overige zijn alleen het gevolg van de interviews. De vraag aan de expert is een oordeel te vellen vanuit haar/zijn expertise. Dit oordeel moet zoveel mogelijk evidence based zijn. Dit kan relevante praktijkervaring zijn of (uiteraard) wetenschappelijk onderzoek in het vakgebied van de expert.

Opslag

Het *werksysteem* verzekert dat er in het lopende onderzoek voldoende, minimaal 8Tb, betrouwbare opslagruimte per onderzoek beschikbaar is om de *relevante data* benodigd om het onderzoek, of delen ervan, te kunnen herhalen op te slaan.

(Betrouwbaar wordt in de interviews benoemd, van belang is vooral dat er backups zijn of andere manieren om terug te kunnen vallen.)

Het *werksysteem* levert functionaliteit om data te *synchroniseren* (in beide richtingen) tussen *onderzoeksdatahouders* en een centrale opslag.

(Centrale opslag kan lokaal TU zijn, maar kan ook de centrale opslag van bijvoorbeeld DropBox zijn. Eén van de interviewers zorgt er voor dat data van meetsystemen automatisch wordt gesynchroniseerd naar een centrale onderzoeksserver, een ander gebruikt DropBox voor dit soort doeleinden).

Het *werksysteem* verzorgt geautomatiseerd een backup van de files en folders door te *synchroniseren* (i.v.m. restores in beide richtingen) met een centrale opslaglocatie.

(Geautomatiseerd betekent vooral dat de onderzoeker hier geen werk aan heeft.)

Het *werksysteem* biedt een opslagplaats waar voor een *relevante periode* de *relevante data* van een onderzoek voldoende snel (sneller dan tape) opgehaald kunnen worden.

Relevante data

Het *werksysteem* biedt de mogelijkheid een vrije, geadviseerde of opgelegde naamgevingformaat en folderstructuur voor de *relevante data* in het lopende onderzoek aan te houden waarbij formats in templates beschreven kunnen worden.

Het *werksysteem* levert een mechanisme om geautomatiseerd versiebeheer op bestanden en data te kunnen toepassen.

(Vooral systemen voor code als Git en Bitbucket zijn populair, maar ook DropBox)

Het *werksysteem* verzekert dat *relevante data* van het lopende onderzoek *vindbaar* en interpreteerbaar zijn door de data gezamenlijk op te slaan, zodanig dat duidelijk is dat ze bij elkaar horen.

Het *werksysteem* maakt het mogelijk data te vernietigen en logt deze activiteit.

(De data zelf zijn niet meer vindbaar, maar het feit dat ze vernietigd zijn, kun je wel terugvinden.)

Het *werksysteem* biedt de mogelijkheid om geautomatiseerd *metadata* aan te maken, bij voorkeur m.b.v. een template.

Interactie

Het *werksysteem* biedt de mogelijkheid tot interactie voor mensen (bijvoorbeeld een userinterface) en bij voorkeur ook machines (bijvoorbeeld een API).

Beveiligen

Het *werksysteem* biedt de mogelijkheid de *relevante data* in het lopende onderzoek te beveiligen.

(Bij beveiligen moet je denken aan vooral authenticatie en autorisatie, maar ook aan encryptie.0

Het *werksysteem* verzekert de integriteit van de data. Dat betekent dat de data accuraat en consistent blijven gedurende hun bestaan.

ELN

Het *werksysteem* biedt de mogelijkheid een *ELN* te gebruiken, waarbij de inhoud van het ELN op verschillende devices beschikbaar gesteld kan worden.

(Vergelijkbaar met *synchroniseren*.)

Het *werksysteem* biedt een *ELN* aan dat de mogelijkheid heeft om o.a. activiteiten en keuzes te loggen.

Het *werksysteem* biedt een *ELN* dat de mogelijkheid heeft om te linken naar relevante data in het lopende onderzoek.

(Denk bij linken bijvoorbeeld aan een hyperlink. Het kan natuurlijk ook een URL zijn of een beschrijving van de locatie.)

Het *werksysteem* biedt een *ELN* aan dat specifiek gemaakt kan worden voor bepaalde vakgebieden.

(Denk bijvoorbeeld aan templates waarin bepaalde processen voor de chemische wetenschap beschreven staan.)

15.3.3. Mail bij de tweede ronde

Beste Expert,

De eerste ronde is voorbij en iedereen heeft commentaar geleverd. In de bijlage vind je twee documenten:
Een worddocument met extra verduidelijking, die op basis van de antwoorden van de experts denk ik nodig was.
Het eerdere Excel document met de scores van alle experts en hun commentaar. Bovendien heb ik mijn reactie gegeven en soms wat voorstellen gedaan.
Ik wil je verzoeken om (in deze volgorde):

1. Het worddocument te lezen;
2. Het commentaar van de andere twee experts in de Excelsheet te lezen te lezen;
3. Mijn commentaar en eventueel voorstel te lezen;
4. Opnieuw een score te geven;
5. Ten slotte op basis van het voorgaande jouw reactie te formuleren, inclusief de reden waarom je eventueel je score hebt aangepast;
6. Waar nodig eventueel een voorstel te doen voor herformulering (zodat jij nog tevredener bent over wat er staat).

Belangrijk is dat je je score geeft op basis van het commentaar van de anderen en de verduidelijking die ik heb gegeven.

Mijn inschatting is dat je een paar tot 10 minuten nodig hebt voor het worddocument, een minuut of 15 voor het doornemen van het commentaar van de anderen en een minuut of 15 om tot nieuwe scores en commentaar te komen. Drie kwartier, maximaal een uur aan werk verwacht ik. Als er net als vorige keer weer iemand heel snel klaar is, hoop ik dat die persoon de bestede tijd deelt, dan kan ik de anderen weer een inschatting geven.

We lopen nu 3 dagen achter op schema op de twee weken die we gepland hadden staan. Ik zou willen voorstellen om op twee rondes van 4 werkdagen uit te komen.
Voor deze tweede ronde is de uiterste inleverdatum dan 16 februari.
Voor de derde ronde wordt dat 22 februari.

Laat het me graag z.s.m. weten als dit schema onhaalbaar voor je is, dan kan ik daar in overleg met de andere experts en/of met jou misschien wat aan doen.

Groeten,
Bert

15.3.4. Mailbijlage bij de tweede ronde

Extra verklaring bij de tweede ronde

Naar aanleiding van het commentaar van alle reviewers, wil ik graag een paar algemene opmerkingen maken over het proces en de resultaten.

Voor mijn afstuderen volg ik de onderzoeksmethode “design science research”. (Bekend gemaakt door Hevner.) Hierin is het de bedoeling in een aantal rondes een construct (ontwerp in de vorm van principes, criteria, een model enz.) op te leveren. Het construct waar ik voor heb gekozen is het werksysteem en in het bijzonder het work system snapshot (WSS). Daar heeft met name Alter weer heel veel onderzoek naar gedaan en dat in diverse publicaties beschreven.

- Het eerste WSS is resultaat van mijn literatuurstudie. Hier kwam een globaal WSS uit gebaseerd op met name resultaten uit US, UK, Canada en Australië.
- Het tweede WSS (dat wat jullie zien) is resultaat van de daaropvolgende 14 semigestructureerde interviews (11 wetenschappers, twee data stewards, die ook zelf onderzoek ebben gedaan en een IT afdelingsondersteuner). Dit is al specifieker voor de TU.
- Het uiteindelijke WSS volgt op de expert review (Delphi methode).

Qua scoping: Het WSS voor mijn onderzoek beschrijft hoe wetenschappers en ondersteuners en (IT) systemen ‘samenwerken’ (een werksysteem vormen) om zo met data in het lopende onderzoek om te gaan dat deze vindbaar en volgbaar zijn.

Het probleem dat we willen oplossen is overigens gebaseerd op interviews die Wim en ik een aantal jaar geleden hielden. Daar bleek dat veel data uit onderzoek op allerlei systemen en datadragers stonden en ze voor mensen binnen en buiten het onderzoek slecht vindbaar en volgbaar waren.

Uit de 14 interviews voor mijn onderzoek bleek dat de initiële probleemstelling waarin werd gesproken over onderzoeksdata te krap was. Behalve de ruwe data heb je vaak ook de metadata en datadocumentatie nodig om delen van het onderzoek over te doen. De onderzoeker is degene die kan bepalen welke data hiervoor nodig zijn en wat dus de relevante data zijn (voor herhaling van onderzoek).

Heel veel literatuur over RDM gaat vooral over gearchiveerde data (bijvoorbeeld in het 4TU datacenter). Gearchiveerde data moeten over het algemeen voldoen aan de FAIR principes om subsidie te kunnen krijgen, dat geldt niet voor data in het lopende onderzoek.

Specifiek voor products and services zegt Alter (vertaald): “Products and services zijn de combinatie van fysieke zaken, informatie en diensten die het werksysteem voor zijn verschillende customers produceert. De products and services van een werksysteem kunnen verschillende vormen aannemen, waaronder fysieke producten, informatieproducten, diensten, immateriële zaken zoals plezier en gemoedsrust en sociale producten zoals regelingen, overeenkomsten en organisaties.”

Alles wat intern of extern door het werksysteem wordt aangeboden of opgeleverd valt onder deze noemer. Ik heb bevindingen uit literatuur en later uit interviews eerst gefilterd voor lopend onderzoek. Daarna heb ik ze gefilterd op vindbaarheid en volgbaarheid. Dus van de 50 criteria die ik vond, vielen er 30 af omdat ze niet voor lopend onderzoek golden en vervolgens viel er nog eens een deel af dat niet direct te maken had met vindbaarheid en volgbaarheid.

Specifiek voor customers en participants geldt dat ze onderdeel uitmaken (per definitie) van het werksysteem zoals Alter het definieert (net als de informatie en technology die ze gebruiken).

Zowel in ronde 1 als 2 beperk ik me tot wat ik heb kunnen vinden en/of met enige mate van zekerheid heb kunnen afleiden. Bijvoorbeeld: Er is in literatuur en interviews niet heel diep ingegaan op het gebruik van een ELN. Het is genoemd en een aantal geïnterviewden heeft verteld wat ze er mee doen in het kader van vindbaarheid en volgbaarheid. Het leidt tot relatief globale uitspraken.

15.3.5. Mail derde ronde

Beste Expert,

Er is consensus over de products and services van het werksysteem. In de derde ronde volgt een laatste check: Is het commentaar van de Expert’s voldoende verwerkt?
Onder de TAB tweede ronde, vind je scores en commentaar van alle drie de experts. Ik heb onder de TAB ‘derde ronde’ de oorspronkelijke stelling gezet, gevolgd door de nieuwe stelling. In sommige gevallen staat er commentaar bij, bijvoorbeeld omdat iemand een vraag stelde, maar soms ook om uit te leggen hoe de formulering tot stand is gekomen. Ernaast staan de scores van ronde 2. Als je een nieuwe score wilt invullen kan dat, zo niet dan kun je een checkje zetten in de kolom: score ongewijzigd sinds ronde 2. Als je de score nog wilt aanpassen, kan dat ernaast met het vriendelijke verzoek een dergelijke wijziging van commentaar te voorzien.

Ik verwacht dat deze ronde de snelste is, aangezien er geen nieuwe informatie is, alleen het commentaar van de andere experts.

Groeten,
Bert

16. Bijlage quotes uit de interviews

ID	Inhoud quote	Codes
1:3	Geïnterviewde: that you publish a paper maybe you made some measurement with some device you get out a couple of nice graphs and maybe you have some raw data format, in whatever file format that is, and then you archive it. The PHD archives it without anything about the metadata without any, like he knows how to read those files, but maybe nobody else will. And two years later you have zero idea and yes it's beautiful that it's archived by the PHD but nobody will be able to use it	archief eisen archief wat
1:5	Geïnterviewde: So that when you say they give the data set that can have several levels? Yeah because if you make a figure that is data right? That's like that level minus what if you say here is a figure. But here is also the actual numerical data that is plotted on the figure. That's an extra level of the data. If you say I'm also providing the raw data and the scripts that I use that's again a difference. So, what you mean when you say it provide people by the dataset.	archief eisen archief wat rd code rd definitie rd levels
1:7	Geïnterviewde: for example think of experimental research. A lot of the times you make twenty-seven failed experiments and you make one successful experiment, from the successful experiment you make a nice plot and the nice figure you publish that. And what about the 27 that failed. That can be very valuable information of course. And a lot of the times it would never make it into the journal.	integriteit rd definitie
1:8	Geïnterviewde: Whereas you could say: by the way if I published a successful one, I also provide the datasets from the unsuccessful ones. And even though I would say that this level of openness, that you really provide the raw data, or the scripts, everything, that's rare. Even today it's yeah there is more emphasis on it.	integriteit rd definitie
1:9	Interviewers: You make it by simulation? Geïnterviewde: Yes.	rd herkomst rd simulatie
1:10	Geïnterviewde: Scripts models to develop all our own scripts	rd code rd herkomst
1:11	Geïnterviewde: They are actual measured patient data. So you could call it experimental data	rd herkomst rd meetdata
1:15	Geïnterviewde: So you can also use some what we call kind of toy problems and you kind of make up models. So you would make up basically a little anatomy where like your spinal cord would be just a cylinder and then a tumor around would be perfect the half cylinder things like that so that then you kind of make up a model yourself.	rd herkomst rd simulatie

1:16	Interviewer: But would you use it for actual research or is it just test something out? Geïnterviewde: We use it for research. For testing first things.	rd simulatie
1:17	Interviewer: OK. And talking about metadata, do you already use metadata in the active phase of your research for your data? Geïnterviewde: I guess you could say try to, so yeah it's some of it is also standardized. So for example for the medical images there is a data standard there's the DICOM standard there. It's a standardized format	md ja md standaard
1:18	Geïnterviewde: That that's more a, if I write software. I try to properly have metadata with it.	md ja
1:19	Interviewer: And what is typically information you would put in metadata like who created it the date it was created..... Geïnterviewde: Yeah, these kind of things also just formats. So for example if a student writes a script and provide the proper readme with it, what is the input how do you call this create. What kind of data ideals is what does that data mean. What are your units, you have to provide the data. This kind of information is just additional metadata for the actual algorithm that they developed.	md definitie md ja
1:20	Geïnterviewde: Yeah I guess so these, Yeah. We do have these. So you would have an author field in the source for a data we do version control on bit bucket for all our software. Yes. We do have that right after project names for at least for the PHDs, they are under projects. So they're I would have to check it for every single piece of code then they write it whether they put in it by the way this is funded by the.	md definitie md ja
1:21	Interviewer: Yeah but from what you just described you don't make this difference I think. Geïnterviewde: Not really I think. Yeah.	md definitie
1:22	Geïnterviewde: It's also a bit like if metadata is only the. Yeah. Who is the author when was this script made what they felt was. He just put in three fields in those three lines in the code and there is your metadata. Interviewer: So you would put it in the code?	md definitie
1:24	Geïnterviewde: So for the software that's all through git and bit bucket and we have just separate branches	del code opsl code
1:25	Geïnterviewde: And for documents and data, basically we use Dropbox	del sync and share opsl share and sync
1:26	Geïnterviewde: And then we have a shared project folder there with the official deliverables and these kind of things.	del centraal opsl centraal

1:27	Geïnterviewde: MMMM, yeah a little bit I guess. So I had the research projects for example in the US over the summer. And then I did some work there. And it wasn't really like me sharing the information it was more a little bit of implementation work of the same code that the implant here in Erasmus so than again that was just through two bit bucket. But I guess if I had to share something then I often just use Dropbox links and then you can set expiry date and you have just a limited access and password protection and all of that	del code del extern del sync and share rd opruimen vlg code vlg ja vlg share
1:28	Interviewers: Students by the way are all using Google Drive. Geïnterviewde: Yeah. Ok Google Drive is the same.	del sync and share opsl share and sync
1:29	Geïnterviewde: Twice it saved my life and basically my computer just went bust and everything was there. Didn't lose a single thing.	bvlg backup
1:30	Geïnterviewde: I do find it's sharing pretty nice actually, especially if you have.....You have to pay, I think like 80 euros a year or something like that. So don't have the very basic, I have one extra but that does allow you for example with the sharing it does allow you to share only for a limited amount of time it doesn't allow you only to share via links but via password protected links. Those are kind of nice to have.	del motivatie del sync and share rd opruimen vlg ja vlg share
1:32	Interviewer: would you like to have some sort of logging of it. Or do you say..... Geïnterviewde: Sure you know any normal nice sharing system would have that as default	logging vlg share
1:33	Geïnterviewde: Not because I would immediately want to police who reads it, but I can totally see why you would want to have that.	logging vlg share
1:34	Interviewer: the data is everywhere. Is that something you recognized it did. Geïnterviewde: Oh yeah definitely yeah.	opsl overal
1:35	Interviewer: So you have the same problem as well you do. Would you be able to collect all the data from one kind of research. Geïnterviewde: I try. So I try to centralize this, that at least my PHDs and my students that they do work on from my Dropbox shared folders. That they their reports there, that they have their figures there. They have the data they generated there. And for the software I just force everybody to use bit bucket and properly commit stuff and keep their research data	fdst groep fdst persoonlijk opsl overal opsl share and sync vind centraal vind folderstructuur vind ja
1:36	Geïnterviewde: But even then it's, I mean you are only one person right? You are not looking behind their shoulders all the time. So it's more trying to set this up from the very beginning and set some kind of policy is that yes I want this and this and this and this and then just try to have people do to here.	ind werkwijze

1:37	Interviewer: It's not something departmental Geïnterviewde: No	ind werkwijze
1:38	Geïnterviewde: You have data on HPC I think? the data gets shared is well or as accessible for more people than .... Geïnterviewde: I think that that is definitely what I would consider active data, because yes students have accounts and they just, you know, do the calculations and then hopefully they would archive and get the final data on Dropbox once it's finished	opsl ext datahouder opsl share and sync rd actief rd definitie
1:39	Geïnterviewde: But that is a difficult issue because we do have students leaving and then their account is still there and their account is no longer accessible and sometimes you have students that left two years ago and they have 10 terabytes of data or hundreds of gigabytes of data and nobody really knows what the student was doing.	bvlg autorisatie rd opruimen
1:40	Geïnterviewde: an issue with HPCs. That it is just kind of separate. So it's not the project drives, it's not the home drives. Like internally it has backup but there is not an easy way to say: if you have your own folder or some sync folder on HPC, then it automatically syncs to your whatever. It is an issue. And personally you can you can install Dropbox on HPC.so you have that. But then again you are kind of syncing everything	bvlg backup opsl ext datahouder opsl share and sync
1:41	Geïnterviewde: You can do selective sync, but then you really have to know what you are doing. You may install a Dropbox account: like you install your Dropbox on HPC then you select from my HPC stuff you select I want to sync this and that... Interviewer: But it sounds like a solution to..... Geïnterviewde: Yeah you're right. I did that! Yeah that's how I did it during my PHD. It is a solution. I'm just saying that is not something that a bachelor student is going to start doing	archit share and sync vind centraal vind ja
1:42	Interviewer: What about data security? So you authorize through Dropbox and on your HPC who gets access to the data? You said before you would like to see who actually access to data if necessary. Geïnterviewde: Yes, sure.	bvlg authenticatie bvlg autorisatie logging
1:46	Interviewer: Do you share encrypted files on your Dropbox? Geïnterviewde: No. It's all the thing we have data is that anonymized at least on it. You know something	bvlg encryptie del sync and share
1:47	Geïnterviewde: Yeah. It's it's a little bit of a double edged sword because one of the main issues in Dropbox is that you share on the other hand certain things you would want to encrypt or at least you would want them to store it encrypted. You know even if it's not encrypted locally, centrally it would be nice if everything is encrypted. And that's not the case with Dropbox. And I think that they don't even offer that service at all so they like even if you have the I know most expensive plans I don't think they encrypt centrally anything. Somehow you would have to encrypt it yourself.	archit share and sync bvlg encryptie del sync and share

1:48	Interviewer: Do you use some sort of versioning on your data sets? Well my imagination is it like you get your raw data and you do some manipulation and you get a resulting set with version 2 or something and then you know maybe more manipulation version 3 and you record whatever you did between the versions…. Geïnterviewde: I guess try to but the problem is I don't really do a lot of the research myself so it's best as I can try to get the students to do it. Naming conventions I definitely have. So I actually have a little naming file, naming conventions file that I share for every students they have project names and dates for naming but for documents I do have version control. So there again I have a Dropbox.	vlg ja vlg transformatie vlg versie werkzaamheden
1:49	Geïnterviewde: No I have like a little file there for just naming conventions for the files for documents. Basically a naming convention for what happens between revisions not as a finalized version etc.. But this is something I kind of just not cooked up myself but check on the Internet and this seemed like a good the reasonable way. But this is mostly for it for documents.	nmc persoonlijk vlg ja vlg naamconventie vlg transformatie
1:51	Geïnterviewde: I think for or for a lot of what we do it's kind of there because again the DICOM format would just keep track of a lot of it. For other things not that much I have to say.	md ja md standaard vlg ja vlg transformatie vlg versie
1:52	Geïnterviewde: Yes I have my. This is a specific typical project folder structure [showing on his laptop] sort of always looks like this. So this is just the students you have. Yeah separately different things and for projects. Yeah kind of a specific folder structure of different aspects of the management Geïnterviewde: you add the file naming convention [text file in the structure] Geïnterviewde: and what it's taught you write financial stuff the personal stuff meetings	fdst groep fdst persoonlijk vind folderstructuur vind ja vlg ja
1:55	Geïnterviewde: So you would imagine kind of like an electronic journal that says: Today I ran this calculation and the input file was this, the output file was that? Geïnterviewde: That could be something. Yeah yeah yeah yeah. And that you store your data somewhere and you can link to it. You give the data of course some unique name.	archit eln dd data bewerken dd eln dd link vind ja vind notebook vlg ja vlg transformatie
1:56	Geïnterviewde: Yeah I totally agree. I absolutely see the point of keeping that kind of journal I don't think you need to .....yeah, good idea. I mean you can always write it down so you can always have some kind of a journal. What did I do today. Yeah exactly like that.	archit eln dd eln dd link vlg ja vlg transformatie

1:57	Geïnterviewde: The linking is difficult because I only have one platform that is an online platform and then there is the file names wherever.	archit eln
1:59	Geïnterviewde: it's the old habits die hard problem but. That that's maybe the biggest challenge I think not that there are not a lot of good software tools, not that there are not good ways of doing things but more that especially in a collaborative environment and you do have a lot of people with different people with different habits. It's not easy to come to an agreement than to actually stick to that. That's the biggest issue I think.	indmot mate van flexibiliteit
1:60	Interviewer: And I think you were already spot on when you mentioned Dropbox. Lots of people are using Dropbox. It's an easy tool. It doesn't have really learning curve. You can just incorporate as you can do whatever you want. Geïnterviewde: And that takes at least a large part of the problem that you don't lose data. You don't have to think about backing it up you. Everything is version controlled. You can go back. I cannot go back forever but at least you can go back I think for months. But in the basic package and it's easy you know. So it's just a habit of doing your stuff on Dropbox and it gets synced. That's already one big step. Not that many people use it. So Yes there are a lot of people using it but there are so if you ask the faculty in this building if I had to guess it's maybe me and maybe one other person from 20 people you know or maybe a couple others	indmot mate van overhead
1:61	Geïnterviewde: You know that's again it's relatively easy to learn. It's easier to keep track of. And you see what is happening they see what is happening. You have a record of what did you discuss in all the meetings. So at least for the small management aspect. It's just easy to.	indmot mate van overhead
2:1	Geïnterviewde: There are people there are a lot of people still who are using external hard disks for data or just their own pc's like the laptop with no backup	bvlg backup opsl ext datahouder
2:2	Geïnterviewde: And there are people who do backups but automatic backups are not as common.	bvlg backup
2:4	Geïnterviewde: So we found out that there are people who are backing up their data but a lot of them are doing it manually.	bvlg backup
2:5	Geïnterviewde: So we do definitely need to increase more awareness for automatic backup.	bvlg backup
2:6	Interviewer: Yeah it's a matter of definition. It's not really a backup. Geïnterviewde: Exactly you are right, it is a copy.	bvlg backup

2:9	Geïnterviewde: So now as a result they even if they use project drive they really use it as a secondary location to do their manual backup so they keep the data on their P.C. and then manually back it up	opsl ext datahouder
2:10	Geïnterviewde: there are a lot of researchers who pay for Dropbox and use Dropbox and they're actually happy with this	opsl share and sync
2:11	Geïnterviewde: Yes of course of course of course. And it does allow till a certain number of version control, so you can go back the same as Google Drive. You can go back to a certain number of versions, you can see what was changed	opsl share and sync vlg ja vlg versie
2:12	Geïnterviewde: People do all sorts of things on pc's, hard drives, paid Dropbox accounts but also free cloud solutions.	opsl share and sync
2:15	Geïnterviewde: in terms of documentation. It's also really varied	ind werkwijze
2:16	Geïnterviewde: So there are people who have handwritten notes or lab journals and there are people who are typing in Word or and there are also some people using OneNote and they seem to be happy with it	dd eln dd papier dd readme ind werkwijze
2:17	Geïnterviewde: And again it really depends on the researcher	ind werkwijze
2:18	Geïnterviewde: One thing I've found out is that with programming many researchers like to use Git.	opsl code
2:20	Interviewer: Maybe some use SPSS it makes the metadata. Other researchers say I have some comments in my scripting and that is kind of the metadata. Geïnterviewde: Indeed. Yeah. So first of all many researchers don't know what metadata is.	ind werkwijze
2:21	Geïnterviewde: So if you are talking of documentation, that was what I was mentioning, people do documents like a readme file even maybe not calling it a readme but similar content indeed so like a lab journal.	dd readme
2:22	Geïnterviewde: One problem with metadata is first of all you need like a good solution that allows you to easily collect the metadata	md vorm
2:23	Geïnterviewde: Another problem is that for example 4TU is using Dublin core and Dublin Core is a very generic committed data standards with the system, the author, the dat created and so on. but it doesn't have any disciplinary specifications and in some fields especially in life sciences there are more and more standards that are being defined but especially in engineering these are very very limited.	md definitie md standaard

2:24	Interviewer: Do you think they would use it if there would be a standard? Geïnterviewde: If it was easy to use if it was offered to them in an easy setting because if they have to do it themselves let's say some writes the code let's say like column headings or something. I don't think they will do that but it would help if there is a template	archit ontwerp indmot mate van overhead md standaard md vorm
2:25	Geïnterviewde: Also it depends I think on the lab leader but also someone from the group for example who is ambitious and happy to do this if they agree to use a template which has these fields such has to be filled in every time.	ind werkwijze nmc groep
2:26	Geïnterviewde: or if there is the program that if it is integrated in the program they are using to already write down their documentation.	dd readme md standaard md vorm
2:27	Geïnterviewde: It could be for example with electronic lab notebooks done you again set up a template and every person has to fill in certain fields.	archit eln md standaard md vorm
2:32	Geïnterviewde: they can use Surf filesender or they can use Surfdrive	del sync and share opsl share and sync
2:33	Geïnterviewde: on my experience what I did is to use emails or to use WeTransfer	del centraal del sync and share
2:34	Interviewer: If they need to put big files they use we transfer or they use the sharing option on the project drive. And what I've heard is people actually going abroad to get a hard drive and fly back with it. Geïnterviewde: Also this happens	del centraal del datahouders del sync and share
2:35	Geïnterviewde: Yeah. So there are a few examples who are doing that but there are also a lot of examples where the supervisor even don't have access. And this is why we try to encourage people to work in a project space. Not everyone has access. It's also much more efficient. But again I think the best working example so far is Dropbox and in some cases a project drive but it has its limitations	del centraal del sync and share opsl centraal opsl share and sync
2:36	Interviewer: Project drive should have the Dropbox functionality in the end? Geïnterviewde: Of course that's what they need.	archit share and sync del centraal del sync and share
2:37	Interviewer: So have you heard about people who need to use encryption? Geïnterviewde: It was more that I was advising you need to use encryption so	bvlg encryptie
2:38	Geïnterviewde: at the folder level or file level you can use VeraCrypt	bvlg encryptie

2:39	Geïnterviewde: but it can be actually easier to encrypt the whole disk	bvlg encryptie
2:40	Geïnterviewde: So for that I have seen a few of times where they had the naming convention so they explained that in the data management plan and I also see one they shared their files with me some of them are following file naming conventions let's say starting with the dates in the year a month day format and so on but	nmc groep
2:41	Geïnterviewde: It really depends on the research group and I had seen one case where they were trying to set up this folder structure and file naming convention so which is very good but that is the only research group so far I have seen that doing that	fdst groep nmc groep vind folderstructuur
2:42	InterviewerIt looks like it's really personal. Is that your experience as well? It seems not really departmental? Geïnterviewde: Yes that's true	ind werkwijze indmot mate van flexibiliteit nmc persoonlijk
2:44	Interviewer: And that includes a folder structure? Geïnterviewde: I think that would be also very good. But again it depends on the department.	fdst groep
2:46	Geïnterviewde: In terms of metadata. So what we explain to them for metadata we say for 4TU it is suiting using Dublin Core but also please think of how you are naming your files how you are structuring your folders and also how you are documenting your files. But the data management plan doesn't specifically ask for this	archief eisen archief hoe archief wat fdst groep nmc groep
2:47	Interviewer: So how did you get your data. Where did you get it from. Geïnterviewde: All sorts of machines.	opsl ext datahouder rd herkomst rd meetdata
2:48	Geïnterviewde: I would either right away transfer to my hard disk or there was a shared drive that I could put the files on and then go to my own P.C. and get the files from there	ind werkwijze
2:49	Geïnterviewde: Then I print that photo and I glue it to my lab journal and then I write and I explain. OK. This is experiment. These are my samples. This is a protocol I've done and this is my outcome.	dd data bewerken dd data verzamelen dd inhoud dd papier ind werkwijze rd eln vlg ja

2:51	Geïnterviewde: So then later on I first started with gluing and then later on I started having everything on Word files so then I would decide I can copy paste let's say if I'm following the same protocol so I copy paste it from the previous experiment the same protocol and then I just digitally get the photo of the protein or DNA gel put it there. Explain what are each samples and what are my outcomes.	dd data bewerken dd inhoud dd papier ind werkwijze rd eln vlg ja
2:52	Geïnterviewde: But this is Word. So if I want I can falsify this so there is no version control. That's one of the difficulties I mean with handwriting I think you can still falsify but at least it is more obvious if you falsify something because you need to scratch it or something like that	integriteit
2:53	Geïnterviewde: We all have different measuring devices in the lab. In a lot of other rooms. And then each time either use the USB stick to get the data, some of them [measuring devices] are connected to a network then you can access that network from your own P.C. and you can download it. Put the data on your P.C. and work on it but it's really experimental research really requires a good working mentality and strategy.	ind werkwijze opsl ext datahouder rd herkomst
2:54	Interviewer: And what did you do with your data on all those different devices when you were done? Did you actually delete it everywhere? Geïnterviewde: And if you want you can delete it. I didn't do it because I wanted to know that I have a backup there in case it is necessary and also in some cases you want the data to stay on that device because now you can line your measuring your sample.	rd opruimen
2:55	Geïnterviewde: We all have a desktop and I stored everything on that desktop. I didn't use a network drive so I use it on the local device and then I made two copies. No I made then just one manual backup to an external hard drive.	bvlg backup opsl ext datahouder
2:56	Geïnterviewde: but the experimental data is really all over the place. Interviewer: So it's all measured data in the end? Geïnterviewde: Yeah.	rd meetdata
2:57	Interviewer: Did you use metadata? Geïnterviewde: No I didn't know what is metadata. So again what I did is I just. In a narrative way I wrote in my lab journal. Interviewer: More like data documentation? Geïnterviewde: Yeah yeah but no metadata.	dd inhoud dd papier dd readme ind werkwijze
2:59	Geïnterviewde: I did always with WeTransfer.	del sync and share
2:60	Geïnterviewde: but if I had to do more few times indeed we had to arrange it with an external device like a USB stick or an external hard drive but most of the time WeTransfer was enough	del datahouders del sync and share

2:61	Interviewer: So how did you keep it secure? How did you make sure that only authorized people could access your data? Geïnterviewde: It was only on my local device. And the hard disks the external hard disk that I had for backup. I locked them in my drawer. That was it.	del datatransfer ind werkwijze
2:64	Geïnterviewde: So then I started to do a year month date and the name of the experiments and I tried to include something in the file name about what it includes so then I know, that was it	nmc persoonlijk vind ja vlg ja vlg naamconventie
2:66	Geïnterviewde: So indeed I did One two three four five. I did that too.	vlg versie
2:69	Interviewer: And how do they do that? Geïnterviewde: I guess that they use a version control system and then they also ask to be properly documented.	vlg ja
2:70	Geïnterviewde: And Dropbox they use for versioning	vlg versie
2:72	Interviewer: From what I've noticed till now is maybe it's something to check with you. researchers are looking for the fastest and easiest way to get their research done. They only want to think about the research and everything around it should just work Geïnterviewde: Exactly, exactly! That's exactly what it is. That's also how I worked. I know also during my PHD and also here because we all have so much pressure that it is not important what you have generated it is only important if you have generated the positive results so that's why they all think let's first get the positive results.	indmot doelgerichtheid indmot mate van flexibiliteit
2:73	Interviewer: So if you do experimental science and you have a few data sets and you don't get a good result from the first four and you get a good one from the fifth what would you do with the first four datasets? Geïnterviewde: I keep it. I've never deleted it because I also don't think it's a good practice to delete it in case again some frauds happens and so on so people can always check what happens.	integriteit rd opruimen
2:74	Interviewer: Will it be in the publication? Geïnterviewde: No. Never	integriteit
2:75	Geïnterviewde: People don't say what didn't work people only report what did work and that's quite tricky. I mean now of course is the discussion of research integrity and I'm not blaming that people are doing bad science. There are a lot of people who are trying to do the good thing but if you publish what didn't work that's not interesting for publishers and as a result publications are always written in a way as if we first did this and we got this very nice result and now we did that. But people don't explain what's happened in between. And that is quite upsetting because it's really fake. I mean fake in the sense that we cut out the parts which didn't work out	integriteit

2:76	Geïnterviewde: Or another thing is to open up the lab notebooks.	integriteit
2:77	Geïnterviewde: Maybe it's not in the paper but then in the additional documents say and the lab notebook which documented the whole research can be found here. That's something we need for transparency.	integriteit
2:78	Geïnterviewde: So regarding that the TU Delft policy says all of the research data, codes, documentation that is necessary to reproduce results, needs to be put in 4TU.	archief eisen archief waar archief wat rd definitie
2:79	Geïnterviewde: If they have like graphs which is a lot of time the case they have like plots showing the results the process data that enabled generation of that figure must be put on 4TU but not their own data. We say it is encouraged but at this moment it is not mandatory requirements because we are also told it's a very big step for researchers and then if we put it that way then no one will do it because they might say no there is no way to do it	archief eisen
2:81	Geïnterviewde: I just want to stress one more time is with experimental research it is important to realize that they rely on so many different measuring devices	rd herkomst rd meetdata
2:82	Geïnterviewde: Data comes from so many different sources and that really introduces an extra level of difficulty to organize everything in a proper way.	rd meetdata
2:84	Geïnterviewde: So it is very good to have some general solutions but there are also many cases where you have to evaluate it on a case level and then try to find a solution because simply, it's just simply impossible to cover all of the needs.	archit ontwerp
2:85	Geïnterviewde: But I haven't heard the specific case where they could restore the data. I can't tell.	bvlg backup
3:1	Geïnterviewde: Biologische data meten we aan de hand van high throughput instrumenten waardoor we gigantische bergen data hebben en die moeten verwerkt worden door analyses. En dan krijg je intermediate data waarvan je moet bijhouden wat je ermee wil doen	rd herkomst rd levels rd meetdata
3:2	Geïnterviewde: Die originele data kan ofwel van eigen meetinstrumenten komen. In dat geval zijn die bijzonder waardevol, kunnen die gevoelig zijn, kunnen die privacy gevoelig zijn, medisch gevoelig zijn. Je noemt het maar.	rd gevoelig rd herkomst rd meetdata

3:3	Geïnterviewde: Of die kunnen van publieke resources komen. En dan maakt het allemaal niet zo uit dan mogen die overal rondslingeren.	rd herkomst
3:4	Geïnterviewde: Of die kunnen van samenwerkingsverbanden komen waar dat de partner eigenlijk bepaalt in hoeverre het gevoelig is.	rd herkomst
3:5	Geïnterviewde: En die hoeveelheden data gaan van enkele tientallen tot honderden giegs tot tientallen tot honderden terabytes. Dus dat is een belangrijk probleem van ons. Hoe breng je dat naar hier? Want 100 terabyte is een grote koffer harddisks die wil je niet over je 10 Gigabit lijntje halen.	del datatransfer opsl capaciteit
3:6	Interviewer: Waar komt die data allemaal vandaan? Van de hele wereld? Geïnterviewde: Dus we hebben niet één bron. Ik werk persoonlijk samen met een groep bij MIT, ik heb in het verleden ook een project samen met surf gedaan om een directe fiber connectie tussen TU Delft en Broad [broad institute] op te zetten.	del datatransfer rd herkomst
3:7	Geïnterviewde: Ik werk ook nauw samen met mensen bij Applied Sciences. Dus daar moet er af en toe data verzet worden. Dat kan in principe binnen TU Delft shares dat kun je gewoon van hier naar ginder kopiëren. Dat is niet zo heel spannend.	del datatransfer rd herkomst
3:8	Geïnterviewde: En dan bedoel shares de bulk, de project store? Geïnterviewde: Vooral de project store,	opsl centraal
3:9	Geïnterviewde: sommige instrumenten zijn aangesloten op laptops en dan moeten we het even van de laptop naar een projectshare transporteren.	opsl ext datahouder
3:10	Geïnterviewde: En buiten data heb je ook documenten, figuren, de Final derivatives die over het algemeen in Dropbox of iets gelijkaardigs leven, code leeft typisch in version control.	opsl code opsl share and sync rd code rd definitie rd levels
3:11	Geïnterviewde: Enkel Github of eigenlijk Gitlab i.v.m. de privacy	opsl code
3:12	Geïnterviewde: Krijg je requirements van je partners dus bijvoorbeeld als je met een Amerikaanse partner werkt, geldt Amerikaanse regelgeving die zegt: Van alle data die gegenereerd is of betrokken is met federal funding die moet je public stellen dus dan moet dat getransferred worden naar een publieke repository moet dat gearchiveerd worden, moet dat publiek opgeslagen worden.	archief eisen archief ja archief waar
3:13	Geïnterviewde: Dan heb je wel binnen mijn vakgebied, heb je publiek gefinancierde instellingen die als enige taak hebben data	archief eisen archief ja archief waar

	opslaan die tientallen petabytes aan data staan van iedereen dat is handig, daar stuur je je data naartoe en klaar	
3:14	Geïnterviewde: Dat betekent dat we in ons projectrapportages zetten dit is gedeposit in een door het vakgebied geaccepteerde repository in de UK of in de VS en niet bij het 4TU datacenter. Bij het 4TU datacenter hebben we de titel gesubmit met een abstract en een verwijzing naar de echte data. En dat is dan voldoende.	archief eisen archief hoe archief ja
3:15	Geïnterviewde: Dus daarin verschillen we misschien van opinie voor mij is data tot het einde van levensduur actief, we zijn behoorlijk actief bezig met lifecycle management op data. En data is pas dood als je de disks delete	rd actief
3:16	Geïnterviewde: de archivering is wat mij betreft een belangrijk onderdeel van je hele lifecycle management van creatie tot use tot reuse. Want als je reproducable science hebt. Dan gebruik je je data en op een moment verliest je data zijn nuttigheid en of dat dat na 1 jaar 2 jaar 5 jaar 10 jaar 100 jaar is weet ik niet	archief eisen archief hoe archief ja rd actief
3:17	Geïnterviewde: Maar ik denk fundamenteel is het gewoon hetzelfde. Het is een onderdeel van hetzelfde proces je data moet ergens staan, je moet ergens access control hebben, je moet beslissen wie verantwoordelijk is voor die data en die namen verschillen misschien in de verschillende periodes maar archiveren is echt wel actieve data	archief eisen archief hoe bvlg authenticatie bvlg autorisatie rd actief
3:18	Geïnterviewde: Dat is onlangs nog gebeurd. Een tijdje geleden, ondertussen al bijna zes maanden, dat was met een laptop maar de snelste manier om van applied sciences 200 meter verder naar hier staat data te transferren is met een laptop op je fiets.	del datatransfer opsl ext datahouder
3:19	Geïnterviewde: Maar als ik data transfers vanuit de VS moet doen, stap ik nog wel eens op het vliegtuig dat is gewoon veel sneller	del datatransfer
3:20	Geïnterviewde: Ehm die data stond bij, ik denk, Amazon en ik denk dat dia gewoon gedumpt is dat die fysiek geshipped is.	del datatransfer
3:21	Geïnterviewde: Vorig jaar nog met een groep in Antwerpen samengewerkt. Waar dat het eenvoudiger is om even op de trein te stappen met een disc of een laptop. Of zelfs met het meetapparaat zo met het apparaat in het begin daar de metingen te doen en dan met het apparaat terug te komen.	del datatransfer opsl ext datahouder
3:26	Interviewer: Metadata in de fase voor archivering, laat ik het zo maar noemen, maak je daar gebruik van? Geïnterviewde: Ja	md ja
3:27	Interviewer: Waar moet ik aan denken wat voor soort.... Geïnterviewde: Een Excel tabel met een unieke identifier met alle kolommetjes met alle metadata die je zou wille	md definitie md ja md vorm

3:28	Interviewer: En die lever je dan apart bij je dataset? Geïnterviewde: Ja die leveren we los	md ja md vorm
3:29	Geïnterviewde: omdat het heel vaak zo is dat de metadata een andere beschermingsniveau heeft dan de genetische	md vorm
3:30	Geïnterviewde: Zeker in het geval dat het gaat om niet humane samples maar dan kan het nog zijn dat de metadata die erbij komt wel gevoelig is dus die die leven volledig gescheiden van elkaar.	md vorm
3:32	Interviewer: Waar komt die unieke identifier vandaan, is dat iets wat jij zelf bedenkt of....? Geïnterviewde: soms is dat gedefinieerd door het lab die de data genereert als het publiek is soms bepalen we die zelf. Soms zijn dat willekeurige getallen als het ware de datum en het uur dat experiment begonnen is.	md vorm
3:33	Interviewer: Het hoeft niet aan een bepaalde standaard te voldoen? Geïnterviewde: Nee, het moet uniek zijn binnen het experiment,	md vorm
3:34	Interviewer: Doe je wel version control op datasets? Geïnterviewde: Niet even strikt als op source code, maar we doen wel data freezes	vlg ja vlg transformatie
3:35	Geïnterviewde: Dus onze primaire data wordt in één keer gegenereerd dus die is gewoon één keer gemaakt, freeze, wordt read only. Niemand komt daar nog aan. Dan heb je primaire analyses die worden meestal één keer uitgevoerd. Soms worden die meerdere keren uitgevoerd. Dan wordt het geparallelliseerd, dus dan heb je parallelle tracks waar dat je verdere analyse op doet.	rd levels vlg ja vlg transformatie
3:36	Geïnterviewde: Maar het is eigenlijk altijd write once dat we werken dus we overschrijven in principe niks en deleten ook niks	rd opruimen
3:37	Interviewer: Zou je van het beginpunt tot aan publicatie. Kun je dan op een bepaalde manier, hou je bij wat er allemaal met de data gebeurt in de tussentijd van versie naar versie of van analyse slag naar analyse slag? Interviewer: Ja, lab notebooks	dd data bewerken dd eln dd inhoud rd eln vlg ja vlg transformatie
3:38	Geïnterviewde: Nu het merendeel van de tools en pipelines die we gebruiken schrijven hun eigen logs van wat er precies gebeurd is	logging md vorm vlg ja vlg transformatie

3:39	Geïnterviewde: Dus wat wordt beschreven in het lab notebook is vandaag heb ik deze analyse gedaan en dan kijk je in die directory en dan zie je van die analyse al die log files en scripts en wat nog staan daaruit kun je reconstrueren wat is nu precies gebeurd.	dd data bewerken dd eln dd inhoud logging rd eln vlg ja vlg transformatie
3:40	Interviewer: En dan kun je verwijzingen aanbrengen naar de dataset bijvoorbeeld? Geïnterviewde: Ja dat zit dus allemaal in de logs.	dd link logging vind ja vind notebook vlg ja vlg transformatie
3:42	Geïnterviewde: Het belangrijkste is dat het Free form is, dat je er mee kan doen wat je wil dat het niet beperkend werkt.	indmot mate van flexibiliteit
3:43	Interviewer: Veiligheid klinkt als een belangrijk onderdeel van de data hier. Wat doe je daar precies allemaal aan? Geïnterviewde: Vertrouwen op de case organisatie dat ze hun boeltje op orde hebben qua intruders. En ik heb gewoon enkel project shares met heel beperkte access lists dus ik weet exact wie bij welke data kan	bvlg autorisatie opsl betrouwbaarheid opsl centraal
3:44	Interviewer: En zou het voor jou belangrijker kunnen zijn om achteraf te kunnen zien wie er allemaal aan de data heeft gezeten? Geïnterviewde: Hmm, Nee de meeste projectshares hebben buiten mezelf 1 persoon. Dus in dat opzicht zie ik dat momenteel niet als een noodzaak. En gezien dat al onze data in principe read only is.....Ik zie niet direct de nood in om al de trails te hebben op onze data.	bvlg authenticatie logging
3:45	Interviewer: Gebruik je encryptie van je data? Geïnterviewde: Nee	bvlg encryptie
3:46	Geïnterviewde: Mostly useless tenzij je heel erg weet wat je aan het doen bent en je de metadata hebt en de metadata zit op een laptop die wel hardware encryptie heeft en die enkel ..... met een remote services, een backup service die ook encrypted is. Interviewer: En die heb je los van de TU geregeld? Geïnterviewde: Ja	bvlg backup bvlg encryptie md definitie
3:47	Interviewer: Is dat een typische gespecialiseerde service? Geïnterviewde: Ja, als een gespecialiseerde Dropbox een secure Dropbox met 100 procent encryptie en wel blablabla. Er komt niemand binnen en dus is het net wat extremer in bescherming.	bvlg backup bvlg encryptie md vorm
3:49	Interviewer: maar de data die op de projectshare staat deel je die wel met bijvoorbeeld zijn collega's in de VS. Geïnterviewde: Ja dat gebeurt en dat is een pijnlijke zaak want Tu Delft heeft voor zover bekend geen faciliteiten om even snel 100 gieg te delen. Dat is een terugkerend probleem.	del datahouders del datatransfer

3:50	Interviewer: En hoe doe je dat nu? Geïnterviewde: Huilen. en hopen dat ik hier nog ergens een disc heb liggen waar dat het op staat want anders mag ik het met 100 megabit op een disk zetten en in een envelop steken en opsturen. Interviewer: Gebeurt dat er ook echt opsturen fysiek? Interviewer: Ja een postpakketje ja. Heb ik hier nog wat liggen? [Zoekt naar harde schijven] Normaal heb ik er 2 of 3.	del datahouders del datatransfer opsl betrouwbaarheid opsl ext datahouder
3:51	Geïnterviewde: De enige optie is dat die partner toevallig ook bij SurfSara is aangesloten en dat die dit via Cartesius [cluster bij Surf] kan doen.	del datatransfer
3:52	Interviewer: En zou het voor jou, Ik snap dat Dropbox te klein is, maar stel dat ze een grote zouden maken, die groot genoeg zou zijn. Zou dat voor jou werken zo'n soort oplossing om data te delen? Geïnterviewde: Misschien Ja, wel af en toe gebruiken we, wat is het Surfdrive? Dat kan tot 250 gig. Ja dus die vind het niet leuk als je er heel grote files opzet.	archit share and sync del sync and share opsl share and sync
3:53	Geïnterviewde: Dus mocht er nu een manier zijn dan dat je gewoon kan markeren, share deze folder even, bij voorkeur vanuit de Unix environment, share deze folder even met deze persoon dan zal ik dat gebruiken.	archit share and sync bvlg autorisatie del motivatie del sync and share vlg share
3:54	Geïnterviewde: Van is er een eenvoudige manier zodat ik gewoon deze folder kan markeren om te delen stel je voor de hele TU storage is een gigantische Dropbox waarbij je kan zeggen dat je deze folder deelt die met die en die mensen.	archit share and sync bvlg autorisatie del sync and share vlg share
3:55	Interviewer: En dan blijft natuurlijk altijd nog dat je 100 terabyte moet transferren. Geïnterviewde: Ja, ja een transfer gaat er altijd blijven.	del datatransfer
3:56	Geïnterviewde: In elk geval de actie om te delen is veel eenvoudiger. Momenteel is zelfs het beslissen van ik ga iets delen dan zit je al heel snel in de logistiek van hoe moet, kan dit logistiek. Als er iemand dan een Dropbox folder met mij deelt dan klik op de link in de mail, zip directory en download daar klaar. Dan gebeurt dat	del sync and share
3:57	Interviewer: We hebben het gehad over waar de data staat, dus eigenlijk bij voorkeur zo veel mogelijk op de projectshare en de metadata bij gespecialiseerde services. Geïnterviewde: Ja en ook op laptops.	opsl centraal opsl ext datahouder vind centraal

3:59	Interviewer: En naamconventies op je files en folders? Doe je dat? Is dat iets persoonlijks of is dat ook iets wat de sectie doet? Geïnterviewde: Dat is niet binnen de sectie, binnen mijn groep probeer ik dat ietwat in de gaten te houden en suggesties te doen. Dat is niet een reglement dus niet echt bindend, als je een goede reden hebt om wat anders te doen kan dat ook.	indmot doelgerichtheid indmot mate van flexibiliteit
3:60	Geïnterviewde: Je moet proberen dingen mogelijk te maken niet alles in hokjes te duwen dus een structuur staat ten dienste van het eenvoudig dingen terugvinden.	fdst persoonlijk indmot mate van flexibiliteit
3:61	Geïnterviewde: Maar als de wetenschap wat anders nodig heeft, dan whatever	indmot doelgerichtheid
3:62	Geïnterviewde: Flexibiliteit is altijd belangrijker dan alle regeltjes volgen vandaar dat het eerder een suggestie is dan een fixed set of rules en er zijn wel een aantal suggesties zo van: doe het op deze manier, daar ga je me binnen twee jaar dankbaar voor zijn	indmot mate van flexibiliteit
3:63	Interviewer: Voor de metadata gebruiken jullie daar een gestandaardiseerde taal voor zoals bijvoorbeeld Dublin Core. Geïnterviewde: Nee, om dezelfde zelfde reden flexibiliteit en functionaliteit boven elk project moet in hetzelfde stramien passen.	ind werkwijze indmot mate van flexibiliteit md standaard
3:64	Geïnterviewde: Dat doet het toch niet en binnen het veld zijn er al twintig jaar mensen die aan het roepen zijn: we moeten een standaard hebben om dit en dat te doen en dus de helft van de onderzoekers die nieuwe dingen aan het doen zijn, wel die past er niet in want daar hadden we nog niet aan gedacht. Dat heb je als je snel nieuwe technologische ontwikkelingen hebt je voorgaande dingen passen niet.	md standaard
3:65	Geïnterviewde: Iedereen gebruikt exact hetzelfde file formaat. Om dit soort data op te slaan Interviewer: En wat voor file format is het? Geïnterviewde: maar je hebt een resum, faststaff, fastqueue, sam, bam, you name it er is een hele list die elk andere soorten informatie bevatten die nodig is	md standaard
3:66	Geïnterviewde: Dus de journals waarin ik publiceer zeggen zet je data bij één van die instituten die gefinancierd is want wij kunnen deze hoeveelheid data niet aan. Dus dat is en dat is ook binnen het veld een vereiste. Grotendeels een vereiste, niet alle journals vereisen dat maar het merendeel tegenwoordig wel dat de dat bij één van de repositories staat	archief eisen archief ja
3:67	Geïnterviewde: Waar het wel nuttig voor is is later als mensen niet enkel reproduceren, maar ook willen verifiëren en ook willen verder bouwen en data integreren in grotere en grotere collecties dat ze toegang hebben tot primaire data en niet tot één of andere afgeleide tabel.	archief eisen archief wat rd levels

3:68	Geïnterviewde: Dat zie je ook wel een afgeleide tabel van: Dit is onze data waarin dat we nog deze regressie gedaan hebben en dan vinden we dat. Primaire data tonen wij hier niet. Nou dan kan je niks meer.	rd levels
3:69	Geïnterviewde: Zeker in het geval dat het gaat om niet humane samples maar dan kan het nog zijn dat de metadata die erbij komt wel gevoelig is dus die die leven volledig gescheiden van elkaar. Er is een unieke identifier die die die linkt. En dus als je enkel de berg data hebt, daar kan je niks mee.	md definitie
5:1	Geïnterviewde: Jazeker we hebben verschillende soorten data. Het zijn geen enorme grote hoeveelheden data, je moet niet zoals in domeinen als astronomie denken aan enorme hoeveelheden data ,maar we doen veel case study's, ben je daar bekend mee?	rd herkomst
5:2	Geïnterviewde: Interviews, het bestuderen van artefacten, websites, portalen voor open data en vooral ook hoe de data nu eigenlijk ontsloten is. Vaak vragen, om te kijken wat er wordt gedaan maar ook benchmarks vergelijken. Hoe wordt er op dit moment gemeten, of het succesvol is dat data wordt geopend.	rd definitie rd herkomst
5:3	Geïnterviewde: Dus het is ook het vergelijken van beleidsrapporten bijvoorbeeld. Ik noemde casestudy's, rapporten maar bijvoorbeeld ook enquêtes, vragenlijsten dat is de grootste hoeveelheid.	rd definitie rd herkomst
5:4	Geïnterviewde: Als je denkt aan de grootte van datasets in mb's, dan zijn de enquêtes het grootst.	opsl capaciteit
5:5	Geïnterviewde: Nee maar ook die ik zelf uitzet, of die mijn studenten uitzetten dus ook mijn afstudeerders. Dat zie ik ook als deels.... het is natuurlijk hun onderzoek maar ik gebruik dat deels ook weer om daarover te publiceren, vaak in samenwerking met de student, ze moeten daar wel toestemming voor geven, want het is hun data.	rd herkomst
5:8	Interviewer: En uit die case study's wat haal je daar uit? Dingen als dagboeken? Geïnterviewde: Nee het is meer het bestuderen van hoe iets in de praktijk aan toe gaat, We nemen vaak een theoretisch raamwerk bijvoorbeeld en kijken naar wat de literatuur zegt m.b.t. een probleem als bij het ontsluiten van data. En dat je dan gaat kijken. Hoe gaat dat er in het echt aan toe?	rd herkomst
5:9	Interviewer: Maar dan is het observatie? Geïnterviewde: Observaties zitten er ook bij	rd herkomst
5:10	Inderdaad vooral interviews binnen de case study's zijn het vooral interviews en observaties. Klopt. Ik had net ook een vragenlijst. We doen ook wel verzameling van cases door middel van ons online onderwijs. Ik ben betrokken bij drie online vakken.	rd herkomst

5:12	Geïnterviewde: dus voor mij is onderzoeksdata vooral vragenlijsten, case studies, interviews.	rd definitie
5:13	Interviewer: Hoe bepaal je wat relevant is van die data en wat niet? Wat bewaar je wel en wat niet? Geïnterviewde: Ik probeer sowieso gezien de GPDR te kijken wat heb ik echt nodig heb voor mijn onderzoek. Wat is de onderzoeksvraag die Ik wil beantwoorden en dus relevant in het kader van een publicatie. Kijk ik met name welke data ik nodig heb om de antwoorden over mijn onderzoeksdoel in de publicatie te kunnen realiseren. Dus als iemand dat onderzoek zou willen reproduceren dan moet hij voldoende data hebben om dat te kunnen doen. Dus ik probeer in ieder geval die data te behouden.	integriteit rd gevoelig
5:14	Interviewer: En waar zet je die data uiteindelijk neer na publicatie? Geïnterviewde: Zo veel mogelijk probeer ik het bij het 4TU centrum for researchdata te zetten als open data. Dat kan niet altijd. Maar als ik interviews heb gedaan bijvoorbeeld dan zet ik het codeboek daar neer omdat die interviews zelf vertrouwelijk zijn en niet openbaar gemaakt kunnen worden maar dan zet ik het codeboek neer zodat mensen snappen hoe kom je aan verschillende codes, hoe heb je dan de analyse gedaan.	ind werkwijze
5:15	Interviewer: Wat gebruik je dan? Geïnterviewde: Zowel Google Documenten als Surfdrive, Dropbox een enkele keer, SharePoint gebruik ik ook maar dat zijn altijd dubbelingen. Dus ik heb het ook gewoon op mijn eigen schijf staan	del centraal del sync and share
5:16	Interviewer: En met wie moet je delen? Geïnterviewde: SharePoint is binnen de case organisatie, maar Dropbox en Google documenten is eigenlijk altijd binnen Europese projecten, dus partners binnen Europa. Interviewer: En ook binnen de EU? Geïnterviewde: Ja altijd binnen de EU	del centraal del extern del intern del sync and share
5:17	Geïnterviewde: Ja ik werk ook wel met mensen buiten de EU samen maar ik geloof niet dat ik daarmee ooit wat gedeeld heb.....Ja, we ik heb er eentje ook met een google spreadsheet. Ik zal er vast wel in het verleden meer hebben gehad maar daar zit geen gevoelige data bij. Wat ik nu bijvoorbeeld doe met iemand uit Taiwan en iemand uit Zwitserland is dat we een literatuurstudie hebben gedaan waarbij we, wat was het, 26 artikelen in detail hebben vergeleken en we controleren elkaars resultaten in een google document, een google spreadsheet.	del extern del sync and share
5:18	Interviewer: Metadata, Maak je daar gebruik van? Geïnterviewde: Ik denk dat het inherent is aan het verzamelen van data, dat je ook metadata hebt. Voor mij klinkt het als: Ja natuurlijk	md ja
5:19	Geïnterviewde: Maar gezien de vraag zal het niet heel natuurlijk zijn voor iedereen. Kijk als je bijvoorbeeld een SPSS bestand maakt op	md vorm

	basis van resultaten verkregen met een enquête dan zit daar altijd je metadata verworven in de labels van die variabele	
5:20	Geïnterviewde: Ik probeer altijd zoveel mogelijk de labels of de beschrijving van de labels te matchen met wat er uiteindelijk in de publicatie staat en ook dus met wat er in de vragenlijst staat zodat iedereen het ook terug kan vinden. Dat is misschien niet altijd het geval maar als je eenmaal gaat publiceren en je gaat die data delen. Dan zul je toch, dan zul je dat duidelijk moeten maken aan mensen wat bij wat hoort welke vragen uit je enquête hoort en bij welke variabele in je SPSS bestand. En hoe heb je dat dan uiteindelijk gebruikt voor je analyses in je papers.	dd inhoud md definitie md vorm
5:21	Geïnterviewde: Ja daar kwam ik achter voor mijn promotieonderzoek probeerde mijn data open te stellen dat ik dacht na. Hier is mijn proefschrift en hier is de data. Succes! Toen was de vraag kun je een readme bestand schrijven over hoe mensen je data kunnen gebruiken. Waarin je die metadata wat beter uitlegt. En toen dacht ik: Ja, eigenlijk is het dus niet helemaal duidelijk. Dus zo'n readme bestand is vaak nodig voor alle onderzoekers die niet in dat project zitten waar jij in hebt gezeten en waar je over vijf jaar ook niet meer alle details van weet.	dd inhoud
5:22	Geïnterviewde: Metadata is niet iets wat de hoogste prioriteit krijgt bij het verzamelen van data. Als je publicaties schrijft dan gaat je aandacht naar de publicatie. En de data ontsluiten, dat doe ik meestal daarna maar wel met een noot in de publicatie naar deze data wordt hier en hier beschikbaar gesteld.	md ja md vorm
5:23	Geïnterviewde: bij 4TU heb je mogelijkheid om eerst een DOI aan te maken en dan later die data ook neer te zetten. Zo kun je zeggen ik wil graag dit plekje reserveren.	archief eisen archief hoe archief ja archief waar
5:24	Geïnterviewde: Maar in de publicatie moet je ook vaak anoniem indienen, je zit altijd een spanningsveld. Mensen vragen soms na de eerste review, waarom heb je je data niet opengesteld? Zelf probeer ik altijd ineen publicatie neer te zetten. De data worden ontsloten via deze link zonder dat bij die link mijn naam is terug te vinden. Dan is het gewoon niet meer anoniem dus is het een beetje lastig.	archief eisen archief hoe archief ja
5:25	Interviewer: En waarom zouden ze jouw naam niet mogen weten? Geïnterviewde: Omdat het blind moet zijn. Het verschilt wel een beetje per journal, sommige journals gaan helemaal over naar een open systeem waarbij de reviewers en auteur elkaar kennen	archief eisen archief hoe integriteit

5:26	Geïnterviewde: Ik weet niet of dat goed werkt en in mijn veld wordt nog helemaal niet gedaan maar dat de reviewers niet weten wie het geschreven heeft dat is heel gebruikelijk, in ieder geval in mijn onderzoeksveld.	archief eisen archief hoe integriteit
5:27	Geïnterviewde: Het kan dat iemand ook vooroordelen heeft, dat jij nog dit publiceert laat ik het maar afwijzen. Interviewer: Of juist andersom Geïnterviewde: Dat kan ook, vriendjespolitiek.	integriteit
5:28	Interviewer: En zo'n readme, waar je het net over had, hoe zou je dat dan noemen? Is dat zoiets als data documentatie? Geïnterviewde: Ja.	dd inhoud
5:29	Interviewer: En die metadata je had het net over SPSS. Maak je ook zelf metadata aan en maak je dan gebruik dan een standaard zoals Dublin Core? Geïnterviewde: Het laatste niet. Ik maak wel zelf metadata aan bijvoorbeeld als ik een SPSS bestand heb en ik zeg die 3 variabelen gebruik ik om een nieuw construct te maken. Dan moet je dat Construct beschrijven waar het uit bestaat, het bestaat uit die variabelen.	md definitie md ja md vorm
5:30	Geïnterviewde: Sowieso als je een dataset hebt op basis van een vragenlijst moet je hem opschonen. Soms zitten er dingen tussen waar je..... deze reacties zijn gewoon helemaal niet bruikbaar of mensen die altijd alleen maar neutral hebben ingevuld bij elke vraag. Daar heb je niks aan. Het kan zijn dat iemand dat echt vindt maar de kans dat het gewoon iemand is die denkt ik vul even snel die lijst in om punten te krijgen voor mijn online vak. Daar zit een risico in. Dat soort antwoorden moet je eruit halen en dat kan iemand niet zien als hij alleen die data heeft dus dat moet je ergens beschrijven hoe dat wel is gedaan. Nou ja dat kan in de publicatie maar vaak is een publicatie er gewoon te kort voor en heb je daar andere manieren nodig om dat verder uit te leggen.	dd data bewerken
5:32	Geïnterviewde: Wat ik dan een voordeel vind bij 4TU in vergelijking tot Dans is dat DANS heel veel open laat dus, ze hebben wel een standaard maar heel veel velden hoef je niet perse in te vullen om verder te kunnen. Dus dan denk je als onderzoeker ook, ik doe dit en dit en dit en dan staat het daar en dan doe ik later de rest, maar dat gebeurt meestal toch niet. Bij 4TU word je gedwongen om die velden in te vullen en als je dat niet gedaan hebt, gaat iemand je een mailtje sturen van goh, we missen deze informatie en zou je die nog kunnen geven want dat vergroot gewoon de herbruikbaarheid.	indmot mate van overhead md vorm
5:33	Geïnterviewde: Ja, Zo gemakkelijk mogelijk en ook zo snel mogelijk. Ja het is toch de werkdruk die veel wetenschappers hebben om zo snel mogelijk dingen te doen. Er zijn heel veel dingen die moeten gebeuren.	indmot mate van overhead

5:34	Interviewer: Alles wat ondersteunend is, moet een zo laag mogelijke leercurve hebben en het moet werken? Geïnterviewde: Ja precies.	indmot mate van overhead
5:35	Geïnterviewde: Ik denk dat dat ook een reden is dat heel veel mensen hun data niet delen, niet openlijk delen, laat ik het zo zeggen. Interviewer: Omdat het extra werk is? Geïnterviewde: Omdat het gewoon extra werk is. Waar op dit moment nog niet zoveel beloning voor tegenover staat.	indmot mate van overhead
5:36	Geïnterviewde: Dat is nou precies de kern van mijn onderzoek. Hoe kun je ervoor zorgen dat onderzoekers wel meer gaan delen en meer elkaars data gaan gebruiken. Eén van de dingen waar ik over nadenk is een combinatie van institutionele en infrastructurele arrangementen.	del motivatie
5:37	Geïnterviewde: dus institutioneel gezien zou je onderzoekers beter willen belonen, bijvoorbeeld tijdens hun jaarlijkse evaluatie. Wil je ervoor zorgen dat zij ook worden geëvalueerd op wat heb je bijgedragen aan delen met jouw community? Wat heb je bijgedragen aan hopen science? Dat dat één van de onderwerpen is, die standaard wordt besproken. Daar word je ook op geëvalueerd.	del motivatie
5:38	Geïnterviewde: Infrastructureel gezien moet je er dan wel voor zorgen dat die onderzoekers daar ook voor op de website van het instituut voor worden beloond. Wat 4TU sterk doet, is zij hebben bijvoorbeeld interviews afgenomen, ook met mij, over wat ik doe of gewoon om een dataset beschrijven. Dan ga je naar de site van 4TU datacenter en zie je daar je hoofd heel groot in beeld staan. Van ja, Jij hebt dit heel goed gedaan. Dat je denkt O ja het is wel erkenning van wat je hebt gedaan.	del motivatie
5:39	Interviewer: Backups maak je die? Doe je dat zelf of vertrouw je erop.... Geïnterviewde: Nee, ik vertrouw er op dat de case organisatie dat doet.	bvlg backup
5:40	Geïnterviewde: wat ik wel doe, is als ik een dataset heb en ik heb er een hele dag aan gewerkt en ik heb ook grote veranderingen gemaakt. Dan stuur ik het voor de zekerheid ook nog eventjes naar mijn mail. Misschien dat ik er morgen aan wil werken maar misschien dat er net even iets gebeurd is in die ene dag dat het toch niet goed is gegaan.	bvlg backup
5:41	Interviewer: Waarom zet je het je dan niet op Surf drive? bijvoorbeeld Geïnterviewde: Zou ook kunnen, gemak denk ik. Het is even een kwestie van delen per mail sturen, maar ook het idee dat je dan als de dag erna gewoon wel goed is gegaan. Je data is niet verloren gegaan.	bvlg backup ind werkwijze opsl centraal

5:42	Geïnterviewde: Dan kun je gewoon een mailtje deleten en dat kan ik gewoon even op mijn telefoon doen. In plaats van dat ik ergens moet gaan inloggen misschien kan dat wel ook met Surfdrive hoor, maar dat weet ik dus niet.	bvlg backup ind werkwijze opsl centraal vind centraal
5:43	Geïnterviewde: Ja precies zo word je er ook eerder aan herinnerd en de mail laat zien, o ja, hier heb ik aan gewerkt en de todo dingen,.....gemak ja.	bvlg backup ind werkwijze opsl centraal
5:44	Interviewer: Heb je al eens een restore nodig gehad? Geïnterviewde: Volgens mij niet, wel een keer van een artikel of iets wat geschreven was maar niet van data. Niet dat ik me kan herinneren	bvlg backup
5:46	Interviewer: En betrouwbaarheid? Wanneer is de opslag betrouwbaar genoeg? Geïnterviewde: Ik denk dat dat in grote mate gerelateerd is of er backups worden gemaakt. Ik vertrouw de case organisatie dat zij dit goed aanpakken in de zin dat er zoveel onderzoekers afhankelijk van zijn	bvlg backup opsl betrouwbaarheid
5:47	Interviewer: En beschikbaarheid? Stel dat het een dag niet beschikbaar is is dat een probleem? Geïnterviewde: Ja is zeker een probleem. Ik heb wel eens als ik thuis werk een probleem met inloggen op weblogin.tudelft.nl en daar word ik niet vrolijk van omdat je gewoon belemmerd wordt in je werk. Beschikbaarheid is wel belangrijk maar dan op een moment dat het mij uitkomt. Als het een dag niet beschikbaar is en ik weet het niet. Dat is natuurlijk geen probleem. Op het moment dat je er aan wil werken moet het er zijn	opsl betrouwbaarheid
5:48	Interviewer: En hoe zorg je ervoor dat alleen geautoriseerde mensen bij jouw data kunnen komen? Geïnterviewde: In principe staat alles op mijn persoonlijke schijf. Dus alleen ik kan daar bij. Op Sharepoint heb ik iets aan data staan waar alleen een paar mensen bij kunnen die betrokken zijn bij de open science hoek. Ik beheer zelf niet de rechten over het autoriseren van mensen maar ik weet dat daar kritisch naar wordt gekeken	bvlg authenticatie bvlg autorisatie vlg share
5:49	Interviewer: En is het wel nodig voor je om encryptie te gebruiken? Geïnterviewde: Tot nu toe niet, ik gooi een variabele er gewoon uit als ik denk het is niet nodig en het is gevoelig, dan haal ik het er gewoon uit. En soms hercategorisering als ik bijvoorbeeld geboortedata heb, dan maak ik categorieën van mensen die in periodes van vijf jaar geboren zijn om het wat minder gevoelig te maken. Om de combinatie van data minder gevoelig te maken.	bvlg encryptie

5:50	Interviewer: Wetransfer heb ik wel eens gebruikt maar niet voor onderzoekdata en presentaties bijvoorbeeld. In de Europese projecten waar ik werkzaam ben geweest waren wij eigenlijk altijd de partner die verantwoordelijk was voor de wetenschappelijke kant zeg en nu schreven we soms wel met andere partners publicaties maar hadden zij geen toegang tot de data. Zij gaven wel hun blik op bepaalde stellingen die we neerzetten bijvoorbeeld maar zij deden niet de analyse van data. Dat deden we altijd zelf dus het is niet nodig geweest om dat te delen buiten de case organisatie.	del sync and share vlg ja vlg share
5:51	Interviewer: Ik kan me ook voorstellen dat als de datasets klein genoeg zijn om naar jezelf te mailen..... Geïnterviewde: Ja dan kun je hem ook naar een ander mailen. Per mail is wel regelmatig gebeurd maar niet ook niet op grote schaal.	del centraal vind centraal
5:52	Geïnterviewde: Wat misschien wel wat gevoelig is maar niet zozeer privacygevoelig waarbij je ook kunt voldoen door bijvoorbeeld een wachtwoord er op te zetten of door het niet per e-mail te versturen maar het alleen te bespreken.	bvlg authenticatie bvlg autorisatie del centraal
5:53	Interviewer: Is het voor jou van belang om te zien wie er allemaal aan de data heeft gezeten, dat je een soort van logging daarvan hebt? Geïnterviewde: Nee dat is niet nodig omdat ik niet met hele grote groepen mensen aan data analyse tegelijk werk	logging
5:54	Geïnterviewde: Wel probeer ik bij mijn eigen promovendi die ik begeleid te zien welke stappen zijn gezet. Wat heb je zelf aan data bewerking gedaan? Maar dan zitten ze gewoon naast mij en kan ik het vragen. Dus het is niet nodig omdat ik niet met grote groepen mensen werk.	logging
5:55	Interviewer: Dat is een mooi bruggetje naar versies. Dus je houdt per stap die je..... per bewerking stap.....dat hou je bij en maak je dan ook andere versies aan per bewerking? Geïnterviewde: Meestal wel, als ik grote stappen maak als ik grote veranderingen aanbreng dan zet ik een nieuwe versie neer. En dat doe ik bijna altijd op basis van datum dus ik zal niet, meestal niet, meerdere versies op een dag maken en anders geldt altijd gewoon met de datum 2019 11, we zitten alweer in 12, 12 02. Als er meerdere versies zijn wordt het A B C. Daarna komt de titel.	nmc persoonlijk vlg ja vlg transformatie vlg versie
5:56	Interviewer: Dat is meteen je naamconventie eigenlijk? Geïnterviewde: Ik zie wel dat heel veel mensen dat niet doen. Die dan een makkelijke titel neerzetten en dan versie V1 of V2 erachter. Maar voor mij werkt dat toch niet zo goed	vind naamconventie vlg naamconventie vlg versie

5:57	Interviewer: Je bedoelt hoe je je mappen structureert bijvoorbeeld? Geïnterviewde: Op basis van, nou ja, ik kan het je laten zien dat is misschien het makkelijkste. Interviewer: En is dat iets persoonlijks of is iets wat je met de hele groep doet?	vind folderstructuur
5:58	Geïnterviewde: Nee dat is wel persoonlijk. Ik weet dat er onderzoek naar is naar hoe je dat zou kunnen doen, maar het is gewoon een beetje gegroeid in mijn praktijk	fdst persoonlijk ind werkwijze nmc persoonlijk
5:59	Interviewer: Het belangrijkste antwoord is eigenlijk, dat het een persoonlijke structuur is. Geïnterviewde: Zeker!	ind werkwijze
5:61	Interviewer: Dus op basis van jouw versioning, zou je dan alle bewerkingen die jouw data hebben ondergaan in de loop van de tijd kunnen volgen? Geïnterviewde: Dat denk ik niet. Deels zit het gewoon in mijn hoofd. Ik weet dat dat niet het goede antwoord is. Een deel,... de belangrijke dingen probeerde ik meteen op te schrijven in een draft van de publicatie die ik maak voor een journal of conferentie. Dus probeer hem wel meteen in een wordbestandje neer te zetten. Een deel daarvan eindigt waarschijnlijk uiteindelijk in het readme bestand. Soms maak ik een apart bestandje met wat aantekeningen voor mezelf met ik heb dit veranderd of dat veranderd, wat je wel terug zou kunnen vinden in is in de e-mail communicatie als je met collega's samenwerkt. Ik heb nu dit veranderd en nu dat veranderd.	ind werkwijze vlg transformatie vlg versie
5:62	Interviewer: Dat is een aardig bruggetje naar lab notebooks. Gebruik je die? Geïnterviewde: Dat is dat word documentje waar ik het net over had. We hebben hier natuurlijk niet echt een lab zoals je bijvoorbeeld bij scheikunde hebt.	dd readme
5:63	Interviewer: Het was ook een beetje misleidend, het is meer een elektronisch notebook dat je bijhoudt. Geïnterviewde: Ik had wel eens bij een collega gezien hoe zij dat deed zij schreef gewoon elke dag op Ik heb dit en dit en dit veranderd, gedaan om gewoon maar bij te houden welke wijzigingen maak je nu. Ik doe het gewoon op basis van gebeurt er iets belangrijks nu of niet. Dan schrijf ik het ergens op.	dd inhoud
5:64	Interviewer: Maar sla je dat weer zo op, dat je het makkelijk terug kan vinden? Geïnterviewde: Dat zit gewoon in die mappen. Het probleem is alleen dat ik soms te veel opsla en dat ik niet meer weet waar ik het moet zoeken. Dat je zoveel bestanden hebt dat je denkt, waar stond het nou ook alweer?	vind folderstructuur vind ja
5:65	Interviewer: Wat je ook heel vaak hoort van mensen is Ik wil mijn competitive advantage, ik kan even niet op het Nederlandse begrip komen, niet weggeven, ik publiceer mijn data pas na publicatie.	del motivatie

	Geïnterviewde: Dat hoor je veel bij open data onderzoek, motivaties om niet te delen.	
5:66	Geïnterviewde: Je ziet dat mensen dat meer hebben als ze zelf heel veel moeite hebben gestoken in de de dataverzameling Ik heb bijvoorbeeld als een case study gedaan bij de Universiteit van Oxford bij de astrophysics afdeling. Die moesten gewoon een voorstel schrijven, dat kost ook moeite, maar je schrijft een voorstel en dan vraag je toestemming aan NASA ESA en weet ik niet wat allemaal, mag ik tijd op de telescoop en hier heb je de data. Dan is het relatief makkelijk en ben je eerder geneigd om dat te delen omdat je het ook van een publieke organisatie hebt gekregen.	del motivatie
5:67	Geïnterviewde: Bij hun is het niet eens zozeer die data, die data genereren ze gewoon met scripts en modellen dus dat kunnen ze zo vaak herhalen als ze willen. Het gaat veel meer om die scripts en die modellen dus de source code die willen ze.....eh... die krijg je niet. Interviewer: Dat snap ik wel. Er zijn goede redenen voor.	del motivatie
6:1	Interviewer: Die actieve onderzoeksdata waar we het net over hebben gehad. Ik heb hem gedefinieerd tot aan het moment van archiveren. For the record. Is dat ook een definitie waar jij mee zou kunnen leven Geïnterviewde: Ja!	rd actief
6:2	Interviewer: wat is voor jou onderzoeksdata? Geïnterviewde: Alle omschrijvingen van sample materiaal dat zijn de questionnaires of surveys die mensen in hebben gevuld, wel anoniem over het grootste gedeelte. Maar ja daar komt ook informatie en data uit rollen. Dat zijn alle notities die ik heb gemaakt over alle procedures in het logboek, in het lab dus, en het algemene lab boek de procedures et cetera die ik moet volgen. Alle fysieke samples eigenlijk ook dat valt er ook onder. Dus dat kunnen dan tanden zelf zijn in buisjes met labels tot op opgeloste tanden in het lab ook in potjes om het zo maar te noemen. Met daar een code op	dd papier rd definitie rd herkomst
6:3	Interviewer: Dus het materiaal is ook data? Geïnterviewde: Materiaal is ook data, ja, ja want aan het materiaal kan je dan weer zien, welke kies het exact is, of daar iets mis mee is of niet, of je dat macroscopisch kan zien.	rd definitie rd herkomst
6:4	Geïnterviewde: Het is ook data het hoort er ook bij. Ik maak ook foto's bijvoorbeeld van tanden, niet altijd, ligt er aan of er iets interessants te zien is of niet. Want anders is het gewoon een kies. Ik heb ook foto's van tanden. Dat is ook data en dan voornamelijk de getallen die dan uit het apparaat komen rollen. Dus dat zijn gewoon Excel sheets met een heleboel getallen.	rd definitie

6:5	Interviewer: Waar komen de data allemaal vandaan? Geïnterviewde: Uit het apparaat bedoel je? Geïnterviewde: Ja, ik hoor meetgegevens, tanden.. Geïnterviewde: Tanden zijn vooral uit Nederland. Het zijn gewoon verstandskiezen van mensen die in Nederland naar de tandarts zijn geweest en daar aan mee wilde doen. Het is alleen maar Nederland, alleen maar moderne tijd. Meetgegevens komen allemaal van 1 apparaat af en, eh, ja is allemaal van 1. Misschien 2 zit ik me te bedenken. Ik heb een deel van de analyses niet zelf gedaan, heeft iemand anders gedaan. Twee apparaten.	rd herkomst
6:6	Interviewer: En metadata maak je daar ook gebruik van? Geïnterviewde: Ja. Maar dan om tanden te omschrijven. De Federal Denton Notation System volgens mij heeft dat een ISO nummer. Dus daar maak ik gebruik van. Ik ben overigens niet.... Niet iedereen doet dat binnen mijn discipline. Ze hebben ook hun eigen omschrijvingen van dingen. Maar ja ik vind, we hebben een systeem dus dat is wat makkelijker	md ja md standaard
6:7	Geïnterviewde: En ja we hebben een soort van standaard soluties die we bij...ja... analyseren met de meetapparatuur. Dus dat is niet zozeer dat er een ISO standaard aan is verbonden, maar dat is wel redelijk standaard. Je weet over welk.. welke solutie mensen het hebben en wat voor soort nummer daar ongeveer aan zit. En je dan er tussen labs met elkaar vergelijken van komt dat dan overeen of niet. Dan weet je in ieder geval de offset daarvan.	md ja
6:9	Interviewer: Maar resultaten zijn toch geen metadata? Geïnterviewde: In dit geval deels. Omdat de meting van een standaard geeft informatie weer over de andere metingen. Dus als die standaard niet op orde is, dan zegt dat iets over de kwaliteit van de rest van je data.	md definitie
6:10	Interviewer: Dus het zijn de omstandigheden van het experiment eigenlijk? Geïnterviewde: Ja, ja	md definitie
6:11	Interviewer: En data documentatie maakte je dat? Geïnterviewde: Ja, ja ik hield dat bij in een Word document en niet op papier. Maar de rest is allemaal in een papieren labnotebook bij mij. Want stel je voor dat je computer crasht	dd papier dd readme
6:12	Want ja met je apparatuur neem je ook allemaal stofjes en dingetjes mee. Je kan dan wel contaminatie veroorzaken, dus volgens mij wilden ze dat uberhaupt er helemaal uit. Dus ja dan zit je daar met je pen en papier	dd papier

6:13	Interviewer: En data delen deed je dat ook? Geïnterviewde: We delen de waarden van die standaarden. Want het is ook relevant voor anderen als de standaarden, als dat een andere waarde geeft. Het is voor lange termijn, voor het functioneren van het apparaat, kan je daar dan wat over zeggen. Dus dat deelden we wel heel erg, dat is dan in een document op de gedeelde computer in het lab, dus daar heeft iedereen ook gewoon toegang toe als je toegang hebt tot het lab.	del motivatie
6:14	Interviewer: Dus dat deelde je met mensen binnen de groep? Geïnterviewde: Ja binnen de groep inderdaad. In principe kunnen ook andere mensen, nee want er zit een slot op, nee alleen maar binnen de groep	del intern
6:15	Geïnterviewde: En je deelt je resultaten als mensen samen op dezelfde publicatie staan, of als je wilt dat iemand anders er naar kijkt, maar dat is minder gestandaardiseerd. Interviewer: En hoe deed je dat dan? Geïnterviewde: E-mail, usb sticks, gewoon de laptop meenemen naar iemand anders en er samen naar kijken maar vooral e-mail.	del centraal del datahouders vlg share
6:16	Interviewer: Is het voor jou belangrijk om te kunnen zien dat de mensen met wie je de data gedeeld hebt, er ook iets mee doen? Geïnterviewde: Ligt aan het doel denk ik. Voor mij was het meer mijn data laten zien, om te kijken of ik iets verkeerds deed of dat zij dachten van: Misschien kan je dit er zo mee doen. Of misschien voor het visualiseren, dat er een andere manier beter was. Ik deel nu data samen met een andere collega om dan samen er een publicatie over te doen	del motivatie logging
6:17	Interviewer: Waar staan al je data? Geïnterviewde: Dat staat op surfdrive en dat staat op.....twee...in ieder geval twee apparaten dus twee computers: een laptop en mijn computer thuis en waarschijnlijk ook nog op een derde op de VU zelf.	opsl ext datahouder opsl share and sync
6:18	Interviewer: En loop je wel eens tegen capaciteitsproblemen op? Geïnterviewde: Nee ik heb echt,.. het zijn Excel sheets. Mijn grootste dataproductie is als ik foto's heb gemaakt van mijn samplesetof van 1 of ander sheet van het apparaat dat loopt dan wat op...	opsl capaciteit
6:19	Interviewer: En wanneer vind je je opslag betrouwbaar genoeg? Geïnterviewde: Als alleen ik er toegang toe heb, of zelf kan bepalen wie er toegang toe heeft. Als het geback-upped is. Als ik er op meerdere locaties bijkan. Dat ik zeker weet dat het op meerdere locaties staat.	bvlg authenticatie bvlg autorisatie bvlg backup opsl betrouwbaarheid
6:20	Interviewer: want dat is kennelijk ook een soort manier om het veilig te maken, dat je het op drie verschillende plekken opslaat. Dat doe je bewust? Geïnterviewde: Ja (lacht) ik ben vrij paranoid..	bvlg backup opsl betrouwbaarheid

6:21	Geïnterviewde: En de data waarmee ik dus wel gepubliceerd heb. Voor zover ik dan de eerste auteur ben, Dat staat nu dus ook op 4TU research data. Dus ja dat staat wel goed.	archief ja archief waar
6:22	Interviewer: En je maakt Backups, heb je ook wel eens een Restore moeten uitvoeren? Geïnterviewde: Nee, nee,	bvlg backup
6:23	Interviewer: En hoe zorg je ervoor dat er alleen geautoriseerde mensen bij je data kunnen? Geïnterviewde: Nou ik maak gewoon vooral gebruik van Surfdrive. Dat zit allemaal achter wachtwoorden of achter een computer met wachtwoorden.	bvlg authenticatie bvlg autorisatie vlg share
6:24	Interviewer: En heb je wel eens encryptie nodig? Geïnterviewde: Ja, ja toch voor de zekerheid. Het is voornamelijk geanonimiseerde data van die questionnaires, maar ja gewoon voor de zekerheid zit daar wel een wachtwoord overheen.	bvlg encryptie
6:25	Interviewer: Over de questionnaires zelf, of over je hele harde schijf? Geïnterviewde: Alleen over de questionnaires	bvlg encryptie
6:26	Interviewer: Is het voor jou van belang om te kunnen zien wie die data dan allemaal hebben gebruikt? Achteraf. Geïnterviewde: Ja, ja het is wel heel leuk om te kunnen zien als iemand dat citeert of... Interviewer: Ik bedoel eigenlijk meer van je hebt ergens opgeslagen en iemand heeft er toegang toe gehad en heeft het bijvoorbeeld gekopieerd of. Geïnterviewde: Gebeurt niet zo heel snel. Als in, iedereen binnen onze groep zijn eigen onderwerp en zijn eigen ding dus het zijn meer... als mensen dingen hergebruiken zijn dat meer ideeën en niet zozeer de data. Er zitten wel duidelijk grenzen wat eigenlijk bij wie hoort. Dus niet dat ik ownership daarover heb, maar we hebben wel redelijk goed begrensd wie wat doet.	logging vlg share
6:27	Interviewer: Naamconventies? Folderstructuren? Geïnterviewde: (lacht) Die gebruik ik nu. Niet zozeer zozeer tijdens mijn PHD. Ik werk wel gestructureerd met folders en ik kan ook wel alles goed vinden, maar dat is meer vanwege een search function dan andere dingen	fdst persoonlijk nmc persoonlijk vind folderstructuur vind ja
6:28	Interviewer: En is dat dan jouw eigen structuur? Geïnterviewde: Dat is mijn eigen structuur.	ind werkwijze
6:29	Interviewer: En versiebeheer, doe je dat? Geïnterviewde: Ja maar dan handmatig dus niet geautomatiseerd. Dus gewoon een versie nummer aan het document.	nmc persoonlijk vlg versie

6:30	Interviewer: Dus waar komt de data vandaan? Geïnterviewde: Dat is eigenlijk net zo verschillend als mijn eigen onderzoek. Als in mensen werken ook met fysieke samples. Of ze dat zien dan staat dat is niet altijd helemaal helder. Ze werken er ook wel mee. Een aantal werken er ook met vragenlijsten of persoonlijke data van mensen. Soms zelfs observaties. En een heleboel meetapparatuur hebben we hier en daar komt heel veel dingen uitrollen Excel een origin files, ja vooral dat. Wat hebben we nog meer? Ik denk dat dat de voornaamste instroom punten van data zijn hier.	rd herkomst rd meetdata
6:31	Interviewer: En metadata, is dat algemeen gebruik? Geïnterviewde: Eh, ik denk, ja.... er wordt wel het één en ander aan gedaan door sommige mensen, maar het is niet een algeheel gebruik, dat mensen dat echt actief bijhouden of actief naar standaarden gaan, voor zover er standaarden zijn binnen hun disciplines.	md definitie md standaard
6:32	Interviewer: En data documentatie? Geïnterviewde: Dat gebeurt, ehh, ligt er ook heel erg aan natruurlijk want sommigen zitten volledig geautomatiseerd en doen alles met Git en dat gaat als een trein	dd readme vlg code vlg ja vlg versie
6:33	Interviewer: En dan heb je het voornamelijk over code natuurlijk? Geïnterviewde: Ja maar ze doen ook een deel van hun labwerk omschrijven zo en dat onder versiebeheer, dus ja er zijn er een aantal die echt heel ver zijn ermee. En dan ook een heleboel papier natuurlijk of niet eens papier maar op de schermen van de laboratoria zelf kladderen. Het is echt heel heel heel breed.	ind werkwijze vlg code vlg ja vlg versie
6:34	Interviewer: En labnotebooks wordt dat gebruikt? Geïnterviewde: We hebben een aantal mensen die op elektronische zitten. Maar ook echt heel veel mensen die gewoon echt papier doen. Ze zijn nu wel zich er meer van bewust dat papier, als daar iets van water en vuur of een Lab Incident bij is, dat ze dat kwijt kunnen raken. Dat het ook wel goed is om een digitale versie of een digitale kopie van dat.....Maar er zijn er nog echt heel veel met papier bezig.	dd eln dd papier ind werkwijze
6:37	Interviewer: Data delen hoe doen ze dat? Geïnterviewde: E-mail.	del centraal
6:38	Geïnterviewde: Ja, ik denk voornaamste e-mail. Mensen zitten nu ook wat meer op de projectdrives. Dat gaat volgens mij echt wel een stuk beter. Ik ken één groep die ook voor externe collaborators echt goed gebruikt en dat gaat ook nog goed. Ook wel eens Surfdrive. Maar ik merk dat mensen toch meer Dropbox of Google zitten, dat ze daar gewoon wat bekender mee zijn. Maar ja, ik denk toch voornamelijk wel e-mail dat mensen dat zo doen. Of gewoon even vragen of mensen erbij komen staan en kijken, maar dat is niet echt delen.	del centraal del sync and share

6:39	Interviewer: Maar hoe deel je dan met iemand van buiten de faculteit, buiten het land desnoods? Geïnterviewde: Je kan nog steeds via e-mail of via links via Surfdrive, of toegang geven tot die projectdrive, of Surffilesender hebben ze, dat kan. Ik weet niet of iedereen daarvan op de hoogte is. Want volgens mij gebruiken mensen dus ook gewoon Wetransfer en dat soort zaken.	del centraal del extern del sync and share
6:40	Interviewer: En denk je dat zij... dat het voor hun belangen.... dat het belangrijk is om te weten met wie ze delen en dat ze achteraf kunnen zien dat die andere ook echt iets met die data heeft gedaan. Gekopieerd, gelezen of wat dan ook. Geïnterviewde: Ligt er denk ik opnieuw weer aan. Als... Vanuit een onderzoeker standpunt is het niet....Als ik er niet direct een benefit van heb. Dan interesseert dat mij ook minder of wat die ander daar mee doet. Op het moment dat zij daar wat mee gaan doen en mij daar compleet niet bij betrekken. Dan word ik wel een beetje chagrijnig, want het is mijn data (benadrukt 'mijn'). Dus dan wil ik daar ook credit voor.	logging
6:41	Geïnterviewde: Het is heel erg persoonlijk, op persoonlijke computers, persoonlijke laptop heel erg Dropbox. Af en toe Googledrive kom ik tegen af en toe Surfdrive. Maar ook projectdrive. Het gaat wel steeds meer die kant uit dat mensen daar wat meer bewust van zijn. Als je daar thuis toegang toe wil hebben, is het gewoon een stuk ingewikkelder dan Dropbox en Google en al die zaken.	opsl centraal opsl ext datahouder opsl share and sync
6:42	Geïnterviewde: Mensen gaan voor de gemakkelijke en gebruiksvriendelijke dingen.	indmot mate van overhead
6:44	Interviewer: Maar spreek je nu voor jezelf of spreek je ook voor de mensen hier? Geïnterviewde: Ook voor de mensen hier. Ik denk, Ik denk gewoon in het algemeen... Ja, op het moment dat je het druk hebt dan heb je er gewoon geen. Als je iets nieuws een nieuwe tool een nieuwe manier van opslaan en dit is echt een gedragverandering die je dan actief zou moeten doen. Ja daar tijd en ruimte voor maken, dat is moeilijk.	indmot mate van overhead
6:45	Interviewer: Hoe denk je dat men in de faculteit met data veiligheid omgaat, dus zorgen dat er alleen geautoriseerde mensen bij kunnen? Geïnterviewde: Ik denk wel heel goed hoor, ik denk naarmate de projecten serieuzer zijn en met externe partners. Ik denk dat daar, dat ze daar erg strak op zijn, en zich er vol van bewust zijn.... Ja, niet, andere mensen moeten daar bij. Ik weet niet hoe sterk ze daar bewust van zijn. Inderdaad ook je laptop meenemen en encrypten.	bvlg autorisatie bvlg encryptie
6:46	Interviewer: En encryptie binnen de faculteit? Geïnterviewde: Gebeurt zeker bij de echt commerciële samenwerkingen waarbij ook een contract, ligt dat nageleefd moet worden. En daar zit het wel goed	bvlg encryptie

6:47	Interviewer: Wordt het ook echt in zo'n contract gevraagd dat je encryptie gebruikt? Geïnterviewde: Ja in ieder geval dat er geen enkele andere..... Ja dat er gararantie moet zijn dat er geen externen bij de data kunnen en dat soort zaken. Dan zijn ze er in ieder geval actiever mee bezig.	bvlg encryptie
6:48	Interviewer: En naamconventies en folder structuren? Geïnterviewde: Nee nee. Nee dat zit er gewoon nog niet echt in. Interviewer: Maar dat is heel persoonlijk denk ik, per wetenschapper. Geïnterviewde: Ja en nee, ik zou het per groep, ik zou het per groep goed vinden. Dat ze in ieder geval een hoofdstructuur hebben waaraan je dan zou kunnen houden en dan misschien wel in submappen, dat je wel gewoon je eigen gang mag gaan, want ja op een gegeven moment wissel je ook weer van Groep, om dan echt alleen maar deze structuur vast te houden. Dat is misschien wat....Lastig omdat...	ind werkwijze nmc groep vind folderstructuur
6:50	Interviewer: En versiebeheer op de data? Geïnterviewde: Zijn ze hier niet heel sterk in. Af en toe zitten er wel mensen tussen die dat helemaal strak georganiseerd hebben, en een aantal groepen zitten ook op GitHub dus dan loopt dat wel beter, maar volgens mij is het inderdaad ook gewoon versiebeheer wat ik ook doe gewoon een getal achter de versie.	vlg versie
6:52	Interviewer: Ik kan me voorstellen dat het bij jou in je lab notebook staat? Geïnterviewde: Nee nee. Zelfs niet als er meerdere versies van mijn datastructuur of mijn publicatie versies zeg maar,.. dat staat niet per versie,...Nou ja, soms staat er: Nu heb ik het geformatteerd naar dit journal, dan staat journal er bij of nu.... Maar. Meestal verander ik van versie als ik echt dingen er in heb aangepast of het format verander of ik heb de bibliografie er in gezet, of ik denk dit hele stuk moet eruit, maar misschien toch niet, dus een andere versie. Maar als ik dat terug moet zoeken moet ik ze even openen.	dd papier
6:53	Geïnterviewde: Dat is wel duidelijk of dat ze ergens een readme file neerzetten van deze versie dit, deze versie dat.... Het zit er niet zo in dat mensen,... ook bij mij niet, moet ik zeggen dat je een readme file ergens neerzet met dit is deze versie, dit is deze versie...Het is heel erg op geheugen	dd readme vlg versie
6:54	Interviewer: Hoe zit het in de faculteit, loop je in de faculteit nog wel eens tegen capaciteitsproblemen aan? Geïnterviewde: Behoorlijk. We hebben een aantal groepen die echt terabytes aan data genereren en dan hebben we hier opzich wel de opslagcapaciteiten, maar ja, ze moeten op een gegeven moment dingen gaan deleten. Wat ga je dan deleten?	opsl capaciteit

6:55	Interviewer: En de betrouwbaarheid van de opslag? Geïnterviewde: Daar is nog wel wat gedoe over geweest want er is hier een aantal jaar geleden is er iets met het netwerk gebeurd en daar hoor ik nog steeds verhalen over, dus dat mensen het liever niet op het netwerk zetten of liever niet alleen op het netwerk zetten want ja: ICT is toch niet helemaal betrouwbaar. En dan vraag je waarom, hoe wat... Dat zit bij sommigen best wel diep.	opsl betrouwbaarheid
6:56	Geïnterviewde: En zo komt het dat hier, een aantal mensen die ik hier heb gesproken hebben dan ook een eigen opslagsysteem dus die waren daar überhaupt nog niet mee bezig of ze zijn later begonnen of dat soort zaken.	ind werkwijze
6:57	Geïnterviewde: automatisch back uppen. Dus ze weten niet dat projectdrive, .. ICT hier doet dat automatisch voor je, dus dat zit dan wel goed. Op zich niks mis mee als ze nog een externe harde schijf hebben of ergens anders, dat is op zich alleen maar goed, maar daar zijn ze ook niet actief mee bezig, ook niet met Surfdrive bijvoorbeeld wat ook automatisch gaat	bvlg backup ind werkwijze opsl ext datahouder
6:58	Interviewer: En ten aanzien van experimenteel onderzoek. Stel je hebt, je doet een experiment vijf keer achter elkaar, de eerste vier keer mislukken en vijfde keer is het raak en kun je publiceren. Wat doe je dan met de data van die eerste vier? Geïnterviewde: Dat ligt er denk ik heel erg aan. Bij mij zelf was het, nou ja als het mislukt dan heb je er gewoon niks aan. Als de standaarden niet goed zijn en de waarden zijn raar en e kan niet corrigeren. Dan is gewoon onzin. Dus dan heeft het ook weinig zin om dat te bewaren. Meestal gooi je weg dan is er wel ergens een record van. Dit is mislukt. Deze redenen. Weg. Maar het meeste,... ja je bewaart het toch maar wel meestal tenzij het echt gruwelijk mis is gegaan. Misschien dat je toch nog, misschien laten corrigeren.	ind werkwijze rd opruimen
6:59	Geïnterviewde: Ik merk hier dat mensen daar ook wel van zijn van dingen weggooien, want ja als het niet bruikbaar is, als de setup niet goed is, of dat ze inderdaad gewoon onzin aan het genereren zijn. Dan gooien ze het ook gewoon weg.	rd opruimen
6:60	Geïnterviewde: Maar ja in principe vind ik dat ook alleen maar beter want anders zit je in die oude versies te kijken. Dan kan je niks meer vinden, dat het gewoon een grote datadump is. Dat zit er wel redelijk in. Ik denk om dezelfde reden dat soms mensen wel dingen bewaren. Stel dat je toch nog kan hergebruiken maar dat ligt ook aan hoe groot die data is. Zeker als het in de terabytes zit. Op gegeven moment moet je ruimte hebben en wordt het toch weggegooid.	rd opruimen
6:64	Geïnterviewde: Maar niet of die vraag van buitenaf kwam. en dat ik daardoor dacht van: oh en nu moet ik het toch maar eens gaan delen. Ja dat gebeurt eigenlijk niet zo heel vaak. Ik heb wel één of	del motivatie

	twee keer de vraag gehad van mensen, waarvan ik wist dat het hun tand was zeg maar.	
6:65	Geïnterviewde: Sterker nog bij de laatste publicatie dat ik erna alle data op het archief had gezet. De peer review zei: Het is geweldig dat je het allemaal zo openbaar zet. Helemaal alsof dat nieuw is zeg maar. Het is ook redelijk nieuw. Het is ook wel nieuw voor onze discipline om dat te doen. Het is niet alsof het genomics is ofzo. Ik meld dit omdat dat één van de pijnpunten is hier. Van waarom zou ik mijn data moeten delen. Niemand vraagt er om. Die komt heel vaak op.	del motivatie
7:1	Interviewer: De eerste vraag is of je dat probleem van die verspreide data ook herkent? Geïnterviewde: Ja, ongetwijfeld	opsl overal
7:3	Geïnterviewde: En dat moet wel op een archive server die een bepaalde data retentie tijd en toegankelijkheid heeft	archief eisen
7:4	Geïnterviewde: Maar in de groep gebruiken wij meestal, de laatste tijd, Zenodo dat is iets die door CERN opgezet is en meer omdat het lower barrier is.	archief eisen archief ja archief waar
7:5	Geïnterviewde: Omdat bij de 4TU dan moet je dan ook allemaal ingewikkelde formen invullen en dan ook nog aan meer eisen voldoen aan je data. Op de ene kant wel goed aan de andere kant is dat vervelend. Dus sowieso is dit archiveren en documentatie van dit hoge niveau geeiste documentatie van je data, ook meer werk voor mensen waar ze niet aan gewend zijn	archief eisen archief ja indmot mate van flexibiliteit
7:6	Geïnterviewde: dus mijn eerste insteek is dan eerst beginnen met iets die niet te moeilijk is. Zodat mensen niet schrikken en als ze een beetje gewend zijn kan je dan meer meer eisen van mensen	indmot mate van overhead
7:7	Geïnterviewde: En dus als je wil mensen stimuleren om met open data te gaan beginnen. Moet je de drempel in het begin zo laag mogelijk maken, net zoals YouTube. Gewoon de eerste tien jaar gratis en dan als iedereen er aan vastzit, ga je allemaal advertenties er door gooien.	indmot mate van overhead
7:8	Interviewer: Hoe bepaal je welke data gearchiveerd moeten worden? Geïnterviewde: Dus in mijn groep hebben wij dat, hebben wij dan voor de afdeling een policy geformuleerd, die staat sinds kort ook op de open science webpage van groep. Daar hebben wij dan, in open data hebben wij twee niveaus gedefinieerd.	archief eisen rd levels

7:9	Geïnterviewde: Een is level Zero en die is heel eenvoudig, gewoon alleen maar punten in de grafiek. Maar dan als een bestand dus dat is processed data en dat stelt niet zo veel voor, voor mensen en stelt ook niet zo veel werk voor voor studenten en voor ons. Maar je hoeft niet te klikken als je met bepaalde software er data uit wilt halen en dat spaart tijd. Het is wel een stap hoger in toegankelijkheid naar de data.	rd levels
7:10	Geïnterviewde: Level 1, dat is dan voor mij de tweede niveau en dat is gewoon met de raw data uit zoals het opgenomen is door de computers die praten met een apparaat en dan samen met alle software en scripts die leidt van dit naar de pdf in de figuur	rd levels
7:16	Interviewer: Als je de documentatie kunt automatiseren, dan wordt het haalbaar. Geïnterviewde: Je kan niet alle documentatie automatiseren zoals: waarom deed ik dit.	archit eln
7:19	Dus wij gaan gewoon door met houtje touwtjes van PowerPoint hier, PowerPoint daar. Wij gebruiken meer online PowerPoint die makkelijk gedeeld kunnen worden. Ik heb mensen die werken met Google sheets, mensen die werken met OneNote. Dus allemaal houtje touwtje en allemaal niet voldoen aan wat ik zou willen.	dd readme ind werkwijze
7:20	Geïnterviewde: ik ben heel erg gehecht aan de Unix filosofie een monolithisch 1 programma dat alles doet, dat gaat nooit lukken.	archit ontwerp
7:21	Geïnterviewde: Dus ik ben meer van een Unix filosofie qua software. Maak je software uit kleine deeltjes die allemaal één ding heel goed doet. Maar dan ook heel makkelijk met de andere stukken kan communiceren zodat jij dan ook flexibiliteit hebt. En dus dat zou zou ik wel graag ontwikkeld zien in een lab notebook.	archit ontwerp
7:22	Geïnterviewde: Dus Electronic lab notebooksoftware waarin alleen maar de basiselementen zitten want per onderzoek, per onderzoeksgroep, zelfs binnen dezelfde afdeling heb je andere manieren van data willen of moeten of kunnen documenteren.	archit eln
7:26	Interviewer: Sommige mensen hebben echt terabytes per seconde en dan wil je misschien even kijken of dat kan. Dat je dat hebt. Interviewer: De vraag is natuurlijk ook of stel dat iemand wat aanvraagt en zijn onderzoek is nog niet goedgekeurd. Wil je het dan ook toekennen dat zijn.... Geïnterviewde: oké. Dat vind ik een beetje bizar. Een researcher, wie is een researcher? Gaan de promovendi dat zelf doen? Interviewer: Dat kan. Geïnterviewde: Dat zou ik niet aanbevelen. Ik zou liever dan... elke promovendus heeft een PI die de begeleider is van een onderzoek. En en en ik vind dat die PI's is zijn degenen die die moet gaan zorgen of. Die. Moet gaan. Of tenminste laat ik het maar zo zeggen ik zie, hoe ik het doe.	opsl capaciteit

	Wij hebben gewoon een plek, wij hebben een teamfolder en daar gaat alles in. En het is logisch dat de PI is degene die dat doet. Die vraagt of. ,Of heeft hij al ruimte toegekend denk ik toch?	
7:29	Geïnterviewde: En je moet wel. Ik ben er ook mee eens. Misschien heb je dan vier terabyte per seconde maar diegene die de data genereert. Ik ga er wel vanuit dat die wordt bewerkt en gedistilleerd naar een hele kleine dataset.	indmot doelgerichtheid opsl capaciteit rd levels
7:34	Geïnterviewde: Je moet het zo zien, je hebt de analoge data, dat is oneindige nauwkeurigheid, oneindige bits rate, oneindige terabytes per second. Maar als je met je instrument en je data analyse enzovoort. Dat is dan een vertaling. Naar een kleine hoeveelheid data. En zo lang als het in principe helemaal duidelijk is wat die processing is geweest kan je op elke punt een grens leggen. En zeggen: dit is mijn ruwe data.	rd levels
7:35	Geïnterviewde: Wij leggen soms de grens in de computer op. Als mijn FPGA gaat voor mij middelen. Dan ben ik die ruwe data van een terabyte per seconde kwijt. Dat is niet zo erg. Omdat ik vertrouw... ik weet precies wat die ding doet. Ik heb het geprogrammeerd. Ik weet wat er gebeurt in die dingen tenminste.	rd levels
7:36	Geïnterviewde: Maar mijn ding heeft geen FPGA en hij streamt gewoon direct naar de harde schijf met een GB elke 10 minuten, maar en dan als ik een script heb en heel duidelijk heb opgegeven en heel duidelijk is wat dat script precies doet. Moet ik dan die GB opslaan? Of is de 100Kb die er uitkomt de enige dat nodig is? Ik ben van mening dat eigenlijk. Dat is ook prima. Wat je verliest is de kans om dat op een andere manier te gaan analyseren. Dus in jouw onderzoek. Als jij dan als een deel van jouw onderzoek is uitzoeken, de invloed van dit staat [deze manier] van analyseren. Als het niet helemaal duidelijk is da t de data op die manier geanalyseerd moet worden en dat daar deels een onderzoek gaat over welke manier is de juiste manier om dat te gaan analyseren dan mag je best wel alleen maar deze data opslaan.	rd levels
7:37	Geïnterviewde: Maar je zou het heel extreem kunnen zeggen en dan heb je alleen maar level 0 data nodig. Geïnterviewde: Omdat je slaat alleen maar op wat in de paper staat en dat wil je ook niet. Dus wat ik denk is dat je moet gaan zoeken naar een pragmatisch middel tussengrond waar op een gegeven moment, ja die data is toch niet meer nodig. Maar als alles heel strak en heel duidelijk gedocumenteerd is van de pad van de apparaat van de sensor die meet naar wat wij opengeslagen hebben, dan mag je wel op elke	rd levels

	stap kiezen. Dit gaan we niet meer opslaan. Omdat anders schiet het ook niet op.	
7:38	Geïnterviewde: De keuze een beetje aan de PI vind ik. Maar kijk maar ik denk er moet wel. Ik ben wel van mening dat ik ben ook wel een persoonlijke mening dat je moet eigenlijk zoveel mogelijk proberen op te slaan.	ind werkwijze rd levels
7:39	Geïnterviewde: We hebben best wel veel harddrive ruimte tegenwoordig. Maar als je het echt over Tb/s hebt dan moet je heel duidelijk een grens stellen waar je niet meer data opslaat. Maar dan is die grens, die moet je proberen zo ver mogelijk naar de ruwe data te gaan duwen. Naar mijn mening tot op het punt dat het niet meer praktisch wordt.	indmot doelgerichtheid opsl capaciteit rd levels
7:41	Interviewer: De andere activiteiten? Dus researcher bedenkt en gebruikt een mappenstructuur? Geïnterviewde: Ja dat doen we.	vind folderstructuur
7:42	Interviewer: Synchroniseert lokale onderzoeksdata en metadata met een centraal aangeboden systeem. Geïnterviewde: Ja dat doen we.	opsl centraal opsl ext datahouder vind centraal
7:44	Interviewer: Werksysteem levert de researcher mogelijkheden voor de aanmaak van metadata, automatisch. Geïnterviewde: Dat doen we een beetje.	md ja md vorm
7:45	Interviewer: Researcher vraagt een PID aan bij het werksysteem (persistent identifier). Dus dat is om je dataset te identificeren. Geïnterviewde: Dat doen we bij Zenodo.	archief hoe archief ja
7:46	Interviewer: Mooi, dus die activiteiten zijn herkenbaar zijn er belangrijke bij die we vergeten zijn? Geïnterviewde: Ik vind dat. Dat is misschien niet zo heel duidelijk maar eh er moet, voor mij wat er moet komen is een lage barrière online systeem om je onderzoek flexibel..... lage barrière systeem om onderzoek gestructureerd te gaan documenteren. Dat mis ik echt erg..... één centrale plek.	archit eln

7:47	Geïnterviewde: één centrale user interface en dan als een toetje als een bonus als die dan gelinkt zou kunnen zijn aan de metadata database. Wat ik in mijn dromen soms ... dan is dat nog mooier maar dat is een droom die gaat in de komende tien jaar niet gerealiseerd worden denk ik. Dus voor mij is dat meer een bonus tenminste een soort van een soort van..... Wij werken met verschillende oplossingen en opties in de groep en geen ervan is ideaal.	archit eln
7:48	Geïnterviewde: Vooral documentatie. Opslag heb ik al voor elkaar min of meer. De. Documentatie en een lab boek. Met een flexibele structuur. Je bent er al bijna met PowerPoint of liever nog google sheets. Behalve dat ze een beetje traag zijn en ik wil de dingen niet opslaan bij Google.	archit eln
7:57	Geïnterviewde: Dat is goeie vraag. Volgens mij is het dan nog steeds een flexibel. Documentatie Systeem. Missen wij, mis ik. Er moet een makkelijk toegankelijk zoveel mogelijk geoptimaliseerde documentatie systeem komen. Je kan niet alles automatisch documenteren	archit eln
7:58	Geïnterviewde: En voor mij is dat ook met dit het geval. Je hebt de documentatie van data uit heel veel verschillende bronnen die op verschillende plekken zit. Dat kan ervoor zorgen dat die documentatie ook linkt terug naar de oorspronkelijke data. Dat is wel een plus.	archit eln dd link ind werkwijze opsl overal vind ja vind notebook
7:59	Geïnterviewde: Dat is ook niet zo anders, omdat dan als jij dat met de hand moet gaan. Alles. Het gaat niet over de werk omdat het werk van documentatie, in principe is 2 seconden maar alles omheen om te zorgen dat je het terug vindt op de juiste plek vindt. Dat is die 10 minuten. En een paar seconden naar 10 minuten dan ben je echt heel veel tijd kwijt als je alles doet. Dus mijn tussenoplossing is en dat zeg ik ook tegen mijn studenten altijd, alles documenteren, een beetje geobsedeerd.	vind ja vind notebook
7:60	Geïnterviewde: Maar dan heel simpel. In mijn software wat ik ooit geschreven heb voor data analyse. Hij zet dan de filename in de window title. Dus als jij dan een screenshot neemt en hem dumpt in een PowerPoint ben je al op 90 procent omdat je hebt de visuele representatie van de data. Je hebt de data name en als die data filename uniek is kan je hem altijd terugzoeken als het moet.	dd readme vind ja vind notebook

7:62	Geïnterviewde: kijk dus wat ik heb gedaan is ik heb hier een shared folder structuur gemaakt waar de ene is dan SteeleLab. En Steelelab, die map heeft dan een hiërarchie van mappen. Waar verschillende dingen inzit. Papers dat zijn papers wat wij nu aan het schijven zijn. Als wij beginnen met schrijven van een paper. Maken wij hier een map in, en zijn alle dus..... we kunnen even gaan kijken. Wat is een goede voorbeeld? [Naam] hier hebben je allemaal de eerste manuscript, daar zijn de figures	vind folderstructuur vind ja
7:63	Geïnterviewde: En dan hebben wij de measurement data en dat is de meeste ingewikkeld. Dit zijn allemaal de systemen waar mijn mensen dan data opneemt dus. Deze zijn allemaal min of meer een main cooler van een grote apparaat met het verschil..... ja apparaten die aan hangen. Die zorgen voor data en bij de Blue Force is een voorbeeld waarbij zelfs drie computers hebben aangesloten, zodat in principe drie mensen kunnen tegelijkertijd kunnen gaan meten. Als je hier in gaat dan heb je...	opsl ext datahouder rd herkomst rd meetdata
7:64	Geïnterviewde: Hij heeft veel gemeten. Dit zijn allemaal experimenten wat hij heeft gedaan. En hij had voor elke sample een unieke identifier. En dan heeft hij dan een naam op basis van die identifier gegeven in die measurment folder. En als hij hier ingaat.	nmc persoonlijk vind ja vind naamconventie
7:65	Interviewer: Die naamgeving laat je hem vrij in toch? Geïnterviewde: Ja in het verleden heb ik het helemaal vrij gelaten. Ik ging er vanuit dat iedereen dat deed, maar ongeveer een jaar geleden bleek dat iedereen,... niet iedereen dat doet.	ind werkwijze nmc persoonlijk
7:66	Geïnterviewde: Of ze geven alleen maar identificaties of namen aan samples die wel gelukt zijn, wil je ook niet.	integriteit
7:68	Geïnterviewde: Omdat iedereen er bij kan. Maar ik laat ze dan in het algemeen best vrij en iedereen heeft zijn eigen manier om dit te gaan organiseren en dat vind ik ook prima	fdst groep nmc persoonlijk
7:69	Geïnterviewde: Ik heb ook persoonlijk een persoonlijke Surfdrive waarop ik referentie letters op bewaar omdat ik gevoelige dingen hier niet opsla.	opsl share and sync
7:71	Geïnterviewde: En wat ik ook zeg tegen iedereen is alles wat heeft te maken met een promotie. Moet....Doe je gewoon een map op computer hier in en gooi het in je project folder.	fdst groep
8:1	Geïnterviewde 2: I think basically all the data involved with basically computational simulations which we are doing	rd actief rd definitie
8:2	Geïnterviewde 3: we do experiments to create active datasets because when you perform your tests you get data runtime data. And I think this idea can be very useful because you need to sometimes filter out the devices that dropped and you don't know what happened.	rd actief rd definitie rd herkomst rd levels rd meetdata

8:3	Geïnterviewde 3: So for me the experimental side of what I was doing here, I was performing a test on a steel bridge, where a vehicle would pass and I would collect information per second. So that is like a huge database. Within three months of terabytes of data and I had one hundred and six sensors installed, which was giving me per second data, different types of data, one was giving me longitudal information one in the vertical direction and one was sheer direction. So that's why we had lots of different information and the numbering and the things that we were getting were very very crucial. Sometimes the device would just stop because the pressure goes off or something happens. But then you need to really filter out the information, otherwise you see suddenly there is a huge strain, so huge that the information doesn't make sense. And then you go back and you check the vehicle just to stop there. So that's why I think this is very important.	md definitie md ja md vorm rd herkomst rd levels rd meetdata vlg ja vlg transformatie
8:4	Geïnterviewde 3: You have to compare data sets in that to see what is really going wrong, so similar things with what you said in experimental and experimental you should be really, you know, taking care about what you already mentioned ,naming your datafiles you know, be really aware what you're running what you know. So we have to have quite a good organization. Otherwise you draw wrong conclusions.	vlg ja
8:5	Geïnterviewde 2: Using logical naming of, let's say, the projects you are running what we're using in finite element simulations. Just having something logically named and then you can always compare. So for example if I'm checking velocity, then I have at least velocity in the name, plus probably the speed was this. And a precise description of what the data tells us that I can later on go back and see.	nmc persoonlijk vind ja vind naamconventie vlg ja vlg naamconventie
8:6	Geïnterviewde 2: most times on the cluster or my Local PC.	opsl ext datahouder
8:7	Interviewer: And do you know always exactly where your data sets are? Geïnterviewde 2: Yes, logical naming of folders	fdst groep vind folderstructuur vind ja
8:8	Interviewer: So you don't need something to find them, you just know where they are? Geïnterviewde 2: Yes basically yes. Not needing some search system to find them.	vind folderstructuur vind ja
8:9	Interviewer: even if you have to go back for a few years. Or is that not common practice? Geïnterviewde 3: It is common practice and that's why indeed in the past you have to have a similar system to really be able find things back after several years. The logical naming of all this is the way I'm arranging it.	fdst groep vind ja vind naamconventie

8:10	Geïnterviewde 3: I can also give you information from 2013 now within seconds because I have a structure right so we we practice this thing because we have to sometimes interact. For example one of our partners asked me to give me datasets from 2016 tests and you know if you don't remember, you will just end up like half an hour never finding it.	del extern fdst groep vind folderstructuur vind ja
8:11	Interviewer: OK so master students that get here, will have to follow the same structure. Geïnterviewde 3: No we do it for them. So I ask them to give me their data and then I will plug it back in the folder. They don't have access to those confidential to replace	bvlg authenticatie bvlg autorisatie fdst groep
8:12	Interviewer: So the directories are confidential? Geïnterviewde 2: Yes some of them were only accessed by people who were involved in the project or had been involved with the project.	bvlg authenticatie bvlg autorisatie
8:15	Interviewer: What kind of security do you need? Is it just authorization that only people authorized can access to data or does it need to be encrypted for instance? Geïnterviewde 2: No. No not really. Geïnterviewde 3: I want to be able to set these shared folders, which you know who is accessing and all that is for us enough to know. Geïnterviewde 2: Sometimes you create passwords. If you can get a password and give it to the correct person by secret.	bvlg authenticatie bvlg autorisatie bvlg encryptie
8:16	Geïnterviewde 2: No about the metadata, but actually for the experiments that we are doing we're collecting lots of different information, different types of information. For example starting from weather also weather was a factor, rain, temperature, degrees centigrade and Mm. Very very different per data set. So we had to create some kind of meta data system but at a very low level and it was both second information. So that kind of thing.	md definitie md ja md vorm vlg ja vlg transformatie
8:17	Geïnterviewde 2: And this is crucial for me and that's what I try to inform you about. If you have this kind of information,n I see a very weird data here for twenty three point five, and then, because my machine is running separately, and you need to know what happened here, did the machine break or is it real value that you are getting. So you need this one and this one should should be coherent.	logging md definitie md ja md vorm vlg ja vlg transformatie
8:18	Interviewer: It does one of you use data documentation that's more like a form where you explain how the data can be read or something. Is that something that's? Geïnterviewde 3: Yes	dd inhoud vlg ja
8:19	Interviewer: And do you use some kind of contracts ,sometimes with people from the outside ,that that gives them access for instance to data or if they want the data you say please sign a nondisclosure agreement or something like that. Geïnterviewde 3: I think with	bvlg contract

	partners we'd do it before starting the project itself. Okay I do not do it for data transaction	
8:20	Interviewer: OK what about data loss.? Are you protected for it protection in terms of like you could do a restore? [..] Geïnterviewde 3: Yeah that one. So when we're doing active data. We create lots of backup. But once we have final set up, generally we don't. Geïnterviewde 3: So when the project is running the we create different times a backup.	bvlg backup
8:21	Geïnterviewde 3: We're trying to get as many intermediate as possible. Yeah. Interviewer: What would be as many as possible is that daily or even hourly ? Geïnterviewde 2: at least whenever I feel like. At least I'm not using certain systems, backup systems or whatever that do these things regularly or on a daily basis or that no no no. For me it's just I have now a valuable set of data so I should make this set secure.	bvlg backup rd levels
8:22	Interviewer: you make a copy over the time? Geïnterviewde 2: Yeah that it can be in week one, week seven or maybe in week twenty four but not regular because that's a waste of resources because sometimes in a few weeks nothing substantial happens and then you're backing up every day a bit. So you have to do wise backups	bvlg backup
8:23	Interviewer: and where do you put your backups Where do you stored them? Geïnterviewde 2: Most times external either on a server or	bvlg backup
8:25	Interviewer: And do you store some kind of information about a version, like this version I did this manipulation for it to get to this version and I did use this manipulation to get to the next version. Geïnterviewde 2: Yeah. in coding. Yes. Geïnterviewde 2: I use indicators to show the change and what will this change entails. Yes.	vlg code vlg ja vlg transformatie
8:27	Geïnterviewde 2: my students do use but not for data. Yeah. Interviewer: So what do they use it for. Geïnterviewde 3: Sometimes they use it. Do they use it as a platform to share and exchange ideas. Geïnterviewde 3: Okay so for example machine learning. One of my students did. They tried to learn from each other and share.	del code del sync and share

8:28	Geïnterviewde 3: when is not important who is important for us when doesn't matter. Geïnterviewde 2: And that's what we have already taken care of by authorizing in advance. Geïnterviewde 3: So logging, time indications. No it's not really an issue.	bvlg authenticatie bvlg autorisatie
8:29	Interviewer: And so sharing you do mostly internally you don't share with people in for instance the US. Geïnterviewde 2: No not at least not on our local drives. No. But we do share with Dropbox and WeTransfer, because if there are external partners, you have to share. Geïnterviewde 3: So we do that but with external facilities.	del centraal del extern del intern del sync and share
8:30	Interviewer: one of the first reasons to start this project was that I talked to people at Applied Sciences and they were saying like: hey in our department data of research could be anywhere, could be on the researchers laptop, on an external drive, on a measuring device, on a dropbox, on a server, it can be everywhere. And he was kind of coordinating he said is a researcher now leaves our departments, I am not able to collect all his data together and say this is from this researcher. Interviewer: Worse than that, even from certain research he couldn't do it because it can be everywhere, is that something that plays for you as well? Geïnterviewde 3: No. Geïnterviewde 2: Well at least I think not everybody's aware of everything everybody knows. Geïnterviewde 2: But if it is Project driven, of course it should be. Everybody should at least be aware, everybody within a project should be aware of what everybody else does.	opsl overal
8:31	Interviewer: So you solve this problem of having it everywhere by making it mandatory. Geïnterviewde 2: Yeah to put it in the structure. Geïnterviewde 3: Yes	fdst groep nmc groep opsl centraal opsl ext datahouder vind folderstructuur vind ja
8:34	Interviewer: So you're not using bulk for instance. K drive some kind. Interviewer: Yeah yeah yeah yeah. But not this intensively, but we could do it sort of because we started not on bulk.	opsl centraal opsl ext datahouder
8:35	Interviewer: Your data storage, for instance your capacity? How do you have enough? And what do you do if you don't have enough Geïnterviewde 2: there were fights in the past. Yes. No. No doubt about it. Geïnterviewde 2: Because there were really limitations everywhere here especially a mailbox this was for years. It was really terrible because if you're doing if you're responsible for a course you get a lot of stuff you know assignments. This that will differ from students.	opsl capaciteit

8:36	Interviewer: If we talk specifically about the place where you put your active data, your research data and we talk about storage capacity and storage reliability. Interviewer: Those are two different subjects but capacity for instance, I can imagine that your data sets are growing over the years. At least that is a normal trend. How do you cope with that? Geïnterviewde 2: Compressing as much as possible and putting more discs in your server. Yeah. Yup. Geïnterviewde 2: Or have alternatives around.	opsl capaciteit
8:37	Interviewer: And what kind of reliability are you looking for. Is it suppose in this folder structure that the underlying hardware crashes, what … Geïnterviewde 2: we rely on this, for sure. Geïnterviewde 3: We depend on ICT. We hope they do the backup and if something goes wrong they can create an image. Geïnterviewde 2: Yes there is trust.	bvlg backup opsl betrouwbaarheid
8:38	Interviewer: So there is a difference probably between the data you really want to have stored and data that you use maybe intermediate or whatever. So relevant data should be stored and who is deciding which data sets are relevant and which are not? Geïnterviewde 3: Yeah project owner. Geïnterviewde 2: So the guy who does the work.	rd levels
8:39	Interviewer: So the searchable sources for you are not so important. The logging of everything that happened, so who has access to a data file when he accessed it. Geïnterviewde 3: when is not important who is important for us when doesn't matter.	logging
9:1	Interviewer: Hoe komen jullie aan je data. Komt dat uit metingen of simulaties? Geïnterviewde: We genereren met onze simulaties behoorlijk wat data.	rd herkomst rd simulatie
9:2	Geïnterviewde: Je kunt het eigenlijk zo zien ik heb een systeem bestaande uit atomen en moleculen en daar maak ik een soort filmpje van hoe dat over tijd of in configuratie space, hoe dat verandert. Als we dat opslaan en soms soms doe je dat dan. Dat gaat over gigabytes tot honderden gigabytes of tot terabytes aan toe mocht het nodig zijn als ze dat niet doen dan is de output vrij, ja eh, de hoeveelheid is vrij klein. We doen wel vaak heel veel simulaties die gescript worden zodat het automatisch gestart en verwerkt wordt. En dat 1000 tot 10000 tot 100.000 zijn. Dus niet hoeveelheid datastorage is het probleem maar het op een goede manier kunnen zoeken.	opsl capaciteit rd herkomst rd levels

9:3	Geïnterviewde: Ja of simulatie condities die dan allemaal resulteren in een hele kleine bestandjes. En iets wat belangrijk is voor mij is dat niet alle data wordt opgeslagen maar meer alle codes scripts enzovoort, zodat als we de data nodig hebben, we het altijd terug kunnen genereren. Interviewer: Want als je het script opnieuw draait, krijg je dan precies dezelfde dataset of zit er een soort van randomizer in? Geïnterviewde: Je krijgt ongeveer hetzelfde maar er zit wel een soort van statistische ruis in.	bvlg backup rd code rd simulatie
9:4	Interviewer: Ik maak onderscheid in mijn onderzoek tussen actieve data en gearchiveerde data dus alle. Mijn veronderstelling is dat alles wat na het onderzoek bijvoorbeeld bij 4TU en Zenodo en dat soort plekken wordt geplaatst. Dat dat archief data is en alles daarvoor wat gedurende het onderzoek wordt gebruikt. Actieve data is. Is dat een redelijke definitie? Geïnterviewde: Ja, dat is het!	rd actief
9:5	Geïnterviewde: En als het onderzoek is afgelopen. Wat we meestal doen is, soms gebruiken we het 4TU datacentrum, maar meestal gebruiken we gewoon support informatie van het tijdschrift. Het tijdschrift vereist ook voor publicatie dat je alle gegevens vermeldt die nodig zijn om simulaties te kunnen uitvoeren en de berekening opnieuw te kunnen doen. Vaak moet je ook gewoon ruwe simulatie resultaten vermelden bij een tijdschrift	archief eisen archief hoe archief ja archief waar integriteit
9:6	Interviewer: En dat is om de validiteit van het onderzoek te kunnen onderzoeken? Geïnterviewde: Ja want het tijdschrift vereist gewoon dat als je iets publiceert dan moet iemand anders met alle gegevens die in het artikel staan kunnen reproduceren wat jij geproduceerd hebt. Interviewer: Is dat dan zo simpel van: Hij heeft de publicatie, hij heeft de dataset en dan moet hij het na kunnen spelen? Geïnterviewde: In feite wel. Dus gegeven dat de algoritmes gewoon open source zijn en gegeven dat die bekend zijn en gegeven hoe het moleculaire model eruitziet moet je gewoon met alle gegevens opnieuw de data kunnen regenereren.	archief eisen archief hoe archief ja integriteit rd code rd definitie
9:7	Interviewer: En gebeurt dat eigenlijk, spelen mensen jullie onderzoek na? Geïnterviewde: nou ja dat gebeurt ook. En het is ook echt absoluut vereist. Dus als je dit niet doet of niet goed genoeg doet, dan krijg je een artikel gewoon terug met het commentaar van de editor je mist dingen. Interviewer: Ah dat doen ze al in de peer review dus?! Geïnterviewde: ja dat is echt de bedoeling. Vroeger was men hier veel minder streng op maar ik heb altijd. Voor mij is het ook belangrijk dat AIO's gewoon hun werk publiceren. en er komt eens iemand met een vraag hoe hebben jullie dit gedaan, dat er precies uitgeschreven staat wat we gedaan hebben.	integriteit

9:8	Interviewer: En die scripts en de parameters die je erin stopt zou je dat onderzoeks data noemen? Of zeg je dat vind ik iets wat ik data documentatie zou noemen? Geïnterviewde: Ja, kwestie van definitie denk ik, ik zou het toch wel onderzoeksdata willen noemen want je hebt het gebruikt bij je onderzoek, dus wat we wel publiceren is alles wat je theoretisch nodig zou moeten hebben om het onderzoek te kunnen herhalen, maar niet bijvoorbeeld echt alle individuele scripts. Dus mensen zouden die dus zelf kunnen schrijven met de informatie uit het artikel.	rd code rd definitie
9:9	Interviewer: En de dingen die je niet publiceert, doe je dat dan een competitieve gedachte? Geïnterviewde: Voor een deel wel van de andere kant heb ik ook gemerkt dat als je je computer code en scripts gaat publiceren op het net gaat zetten. Binnen de kortste keren krijg je vijf verzoeken per dag van Chinezen, Vietnamezen en Pakistanen die het ook willen gebruiken nergens verstand van hebben en zeggen van ja het werkt niet. Je komt om in de vragen. Het andere deel is concurrentie	del motivatie indmot doelgerichtheid
9:10	Interviewer: En die code. Wat doe je daarmee? Doe je dat in github systeem? Geïnterviewde: Ja vaak wel en dan gewoon gesloten voor de afdeling of je onderzoeksgroep. Interviewer: En Iedereen in de onderzoeksgroep kan er weer wel bij? Geïnterviewde: Ja in principe wel.	opsl code
9:11	Interviewer: En gebruik je ook de versioning component van Github? Geïnterviewde: Je bedoelt dat je verschillende versies hebt en versie management van de code? Ja, ja.	vlg code vlg ja vlg transformatie vlg versie
9:13	Interviewer: En doe je dan nog iets met een naamgeving of een opslag structuur. Ik heb bij andere afdelingen gezien dat ze een bepaalde folderstructuur voorschrijven voor de hele afdeling en dat ze willen dat de dataset altijd een bepaalde voorloper heeft met een datum of iets dergelijks. Geïnterviewde: Nee absoluut niet. Het probleem is ook dat het toe dermate eh, nou kijk als dingen gestandaardiseerd zijn, dan kun je dat makkelijk doen maar de ene keer is het dit en de andere keer is het dat en moeten we dit formaat gebruiken, de andere keer dat formaat. Dat is geen uniform formaat en dus geen uniform dingen die we willen opslaan dus als we structuur willen aanbrengen dan kost dat ons heel veel tijd.	fdst groep fdst persoonlijk indmot mate van flexibiliteit nmc groep nmc persoonlijk
9:14	Interviewer: Iedereen bepaalt dus zelf hoe die dat eh..... Geïnterviewde: Ja voor elk onderzoek en voor elk type artikel is dat weer anders er is geen uniek formaat.	fdst persoonlijk nmc persoonlijk

9:15	Interviewer: Heb je dan wel een eenduidig formaat voor allemaal. Per onderzoek als je met meerdere onderzoekers werkt bijvoorbeeld of zeg je van iedereen doet dus gewoon voor zichzelf Geïnterviewde: Nou ja kijk, de data komt natuurlijk op een manier uit het programma. Dat programma is voor iedereen dezelfde maar voor vrijwel iedereen hetzelfde. Kijk als je die data dan weer bewerkt, nou ja... eh, de data structuur is ook dermate simpel deze variabele is dit en deze is dat en zo heb ik er een stuk of tien. Meer is het niet. We hebben niet het issue of zo dat we datasets met honderden parameters met elkaar moeten vergelijken. De dataset die we in het begin genereren is dus heel groot. Daar maken we een hele kleine dataset van. Gewoon een paar getallen. Dus deze vijf grootheden zijn deze waarde plus deze meet onzekerheid. Dan is het gewoon voor iedereen meteen duidelijk wat het is dan heeft het aanbrengen van een onderliggende datastructuur dat altijd op zo'n manier opslaan. Dat is in feite overkill	vlg ja
9:16	Geïnterviewde: Er zijn binnen mijn vakgebied ook pogingen gedaan om het file output formaat te standaardiseren en om input te standaardiseren. Dat is dus totaal niet opgepakt door het vakgebied. Mensen willen daar gewoon geen tijd aan besteden omdat het dermate rigide is en het kost gewoon heel veel tijd om als je iets nieuws wil doen om binnen voorwaarden te gaan. Iedereen wil iets nieuws doen op een andere manier. Dat wordt hem gewoon niet.	indmot mate van flexibiliteit md standaard
9:17	Interviewer: Dat heb ik ook meer gemerkt. Het is toch de tools gebruiken met een zo laag mogelijke leercurve, het moet gewoon werken Geïnterviewde: Ja precies want de tools zijn ook dermate ingewikkeld dat ik er liever minder aandacht besteed aan de tools dan hoe ga ik de data verwerken. Want dat is het wetenschappelijke probleem ook niet.	indmot mate van overhead
9:18	Interviewer: En data over de data? Daarnet had je het over die dataset die uit tientallen parameters bestaat, heb je bijvoorbeeld ook metadata of data documentatie die beschrijft wat die parameters dan inhouden of iets dat voor iedereen evident? Geïnterviewde: Vaak is dat gewoon evident. De metadata kan heel kort.	md ja
9:19	Interviewer: Waar moet ik dan aan denken. De metadata? Wat beschrijven jullie in de metadata? Geïnterviewde: Nou ja wat in de data staat is meestal vrij kort, vaak 1 zin. Vaak een verwijzing naar het artikel: Die parameter van de vergelijking staat in de eerste kolom, een andere kolom is die parameter uit die vergelijking. Interviewer: En dan is je publicatie in feite je metadata. Geïnterviewde: In feite wel ja, Want daar staat het dan uitgelegd.	md definitie md ja md vorm

9:20	Interviewer: Bijvoorbeeld wanneer het is gemaakt, door wie het is gemaakt, op welk systeem het is gemaakt. zijn zijn dat waardes die belangrijk voor je zijn? Geïnterviewde: Nee de versie van software is wel een belangrijke maar met welke een computer, met welke compiler, met welk operating system of wat de stand van de maan was, dat houden we niet bij.	md ja
9:21	Interviewer: En gebruiken jullie lab notebooks? Geïnterviewde: Nee eigenlijk niet. Geïnterviewde: Nee eigenlijk niet, is gewoon te veel werk. Interviewer: En Lab journals? Geïnterviewde: Sommigen houden dat zelf bij, wat ze doen en waar wat staat maar niets dat er een soort dag tot dag te vinden is wat iedereen gedaan en waar de files staan enzovoort. Van de andere kant je kunt, als je gewoon een directorystructuur hebt met al je folders. Je kunt dan heel makkelijk zoeken op datum en op file en op titel.	fdst persoonlijk nmc persoonlijk vind folderstructuur
9:22	Interviewer: Maar dan gebruik je dus wel een directorystructuur? Geïnterviewde: Ja dat wel, dus laatst bijvoorbeeld had ik een probleem, ik wilde een algoritme hergebruiken. Ik heb het ooit eens een keer geprogrammeerd en dat was in 2001, 18 jaar geleden en ik kon daadwerkelijk op mijn computer in een minuut of 10 mijn originele source code van 18 jaar geleden terugvinden.	fdst persoonlijk vind folderstructuur vind ja
9:23	Geïnterviewde: En wat de AIO's doen, de belangrijkste software en de belangrijkste scripts, die zet ik er gewoon bij zodat ik er ook in kan zoeken.	ind werkwijze vind ja
9:24	Interviewer: Waar staat het allemaal? Allemaal lokaal op je Mac? Geïnterviewde: Ja en ik heb een back up zo'n cloud ding en ik ben het ook op een harde schijf. Waarvan er in principe eentje hier ligt en eentje thuis, dus ik heb in principe vier kopien. Interviewer: En wat gebruik je nou aan cloud opslag? Geïnterviewde: Cloud? Interviewer: Ja je zegt ik heb zo'n cloud ding. Geïnterviewde: Nee ik heb een Time Machine van Apple, die staat hier, dat is een een doosje	bvlg backup opsl ext datahouder
9:25	Interviewer: Je gebruikt iets als Dropbox of Google drive? Geïnterviewde: Nee, ik heb een gruwelijke hekel aan Dropbox.	del sync and share
9:26	Interviewer: Hoe doe je dat dan met data delen bijvoorbeeld? Geïnterviewde: E-mail omdat het klein is. Geïnterviewde: Als het echt grote dingen zijn dan is het Surf Transfer van Surf Sara, dat is wat veiliger dan WeTransfer.	del centraal del sync and share
9:27	Interviewer: Want dat is belangrijk voor jullie, veiligheid? Geïnterviewde: Hmmm, mja, Het is makkelijk Surf Transfer werkt wel aardig en WeTransfer is toch commercieel en je weet toch niet wat mensen er mee doen. We gebruiken ook wel eens data waar namen in staan en men heeft wel eens gezegd, gebruik liever Surf Transfer dan WeTransfer. Nou ja, doen we dat	del motivatie del sync and share

9:29	Interviewer: Dus de data handling, als ik het even heel groot en vaag maak, is iets wat heel persoonlijk is bij jullie? Iedereen mag dat voor zichzelf bepalen hoe die dat doet? Geïnterviewde: Juist omdat alle onderzoeken zo verschillend zijn. Juist om die reden omdat iedereen andere dingen doet. Als we dat willen uniformeren zijn we er zo 5 jaar aan kwijt en een dag in de week om het up to date te houden en daar hebben we de tijd niet voor.	fdst persoonlijk indmot mate van flexibiliteit nmc persoonlijk
9:30	Interviewer: En slaan jullie ook dingen op op de TU systemen of is het daadwerkelijk alleen eigen systemen? Geïnterviewde: Ja eigen systemen, we hebben ook een cluster met 1000 cores en een raid systeem en daar hebben we ook een backup van.	opsl ext datahouder
9:31	Interviewer: Nou backups maak je dus, moet je wel eens een restore uitvoeren? Geïnterviewde: Nee nooit gebeurd	bvlg backup
9:32	Interviewer: Goed eigenlijk. Loop je wel eens tegen capaciteitsprobleem aan? Geïnterviewde: Nee, valt wel mee omdat ik gewoon hele grote.... De echte hele grote files die data die datasets die allemaal bewerkt zijn en eh als ze...Als dat klaar is gooien ze weg en dan werken we gewoon met de gereviseerde dataset. Plus het feit dat de code en alle parameters om over te doen.... we een grote dataset altijd opnieuw kunnen genereren. Ik denk dat ik in de 22 jaar dat ik met onderzoek bezig ben, dat ik eh... ja de de echte belangrijke data dat is minder dan een terabyte. Dat is iets wat je gewoon op een normale Mac kan opslaan.	opsl capaciteit rd opruimen
9:33	Interviewer: En die grote datasets blijven gewoon op hpc staan bijvoorbeeld? Geïnterviewde: Ja, tot we een keer uit ruimtegrebruik wat weg moeten gooien, nou dan gooien we het weg. We kunnen niet alles bewaren.	opsl capaciteit rd opruimen
9:34	Interviewer: En iets van encryptie ofzo gebruik je encryptie dat je je data versleutelt bijvoorbeeld voordat je het met iemand deelt? Omdat bijvoorbeeld er namen in staan wat je wil beveiligen. Geïnterviewde: Nou ja, mijn harddrive en mijn backup zijn allemaal encrypted. En ook data opslag op het cluster. Interviewer: Data opslag op het cluster is encrypted? Geïnterviewde: Volgens mij wel. Interviewer: Ja ik heb ook een linux pc en die is encrypted en mijn mac is encrypted en al mijn opslag op harddrives is encrypted.	bvlg encryptie
9:35	Geïnterviewde: Even kijken, even nadenken. Wat ook een beetje een ding is, dat 4TU opslagsysteem wordt mooi gepromoot. Op zich vind ik het ook wel leuk en ik heb er ook een aantal dingen geponeerd, maar er zitten wel een aantal grote nadelen aan het systeem vast voor mij. Boven de gigabyte moet je gaan betalen. Het andere punt is dat ze het maar voor 10 tot 15 jaar opslaan.	bvlg autorisatie

9:36	Geïnterviewde: Terwijl je kunt het opslaan bij tijdschriften en dan staat het eeuwig. Bij een tijdschrift is het gewoon onderdeel van de publicatie. Die publicatie is er voor altijd. Dus je kunt ook tijdschriften die uit de negentiende en de achttiende eeuw zijn, nog steeds gewoon te vinden. Die zijn er nog, en de data die daarbij hoort, die staat gewoon op papier dus die is er nog. Dus als je kijkt hoe lang een academische carrière duurt, die duurt 35 tot 40 jaar van het moment dat je AIO bent tot het moment dat je met pensioen gaat. Iemand is met zijn masters op zijn 23ste klaar, pensioenleeftijd is 68, dus je carrière is 45 jaar, dus dan zou ik zeggen dat je voor een halve eeuw of eeuw moet opslaan en niet voor 10, 15 jaar.	archief eisen archief ja
9:37	Interviewer: En bij zo'n tijdschrift krijgt een dataset ook een identifier bijvoorbeeld? Geïnterviewde: Dat is dan de identifier van het artikel en die behoort dan bij..... Interviewer: DOI Geïnterviewde: Ja. je krijgt dan een DOI van het tijdschrift. Ik heb ook wat dingen staan bij 4TU met DOI en ik ben benieuwd hoe dat uitpakt, wat er over tien tot vijftien jaar gebeurt, of ze een e-mail sturen met wil je het nog behouden of dat ze het gewoon deleten	archief eisen archief hoe
10:1	Interviewer: Dus het staat in een Dropbox het staat op een laptop het staat op een harde schijf op een meetsysteem enzovoort. Eén van die mannen coördineerde het IT werk op de afdeling die zegt dat op het moment dat een onderzoeker hier wegloopt, ben ik niet in staat om al zijn data bij elkaar te vegen. Op het moment dat iemand aan mij vraagt kun je de data voor dit onderzoek bij elkaar vegen, dan kan ik dat eigenlijk ook niet. Geïnterviewde: Heel herkenbaar	opsl overal opsl share and sync
10:4	Interviewer: Inderdaad de actieve Onderzoeksdata. Ik heb dat gedefinieerd als eigenlijk alle data die je hebt tot publicatie. Tot nu toe is dat een definitie waar iedereen mee kan leven. Geïnterviewde: Denk ook niet dat het heel anders kan.	rd actief
10:5	Geïnterviewde: In feite is het, om terug te gaan naar je vraag, dat je twee soorten onderzoeksdata hebt. Enerzijds alles wat nodig is om dat model, dat simulatie model te kunnen runnen. Dat is één ding. Vervolgens gegeven het feit dat je dat gerunde simulatiemodel hebt, genereer je een bak met data. En beiden zijn in feite onderzoek data waarbij vaak de eerste set van data daar heb ik maar beperkt controle over Ik krijg vaak modellen van anderen met allerlei clausules over de over de data die daar aan ten grondslag ligt dat je niet mag delen etcetera.	bvlg contract rd definitie rd gevoelig rd herkomst
10:6	Interviewer: Zou je kunnen samenvatten dat de inputdata altijd meetgegevens zijn? Geïnterviewde: Nee, ze kunnen ook uit andere modellen zijn gemaakt. Geïnterviewde: Klimaatdata is andere modellen. De resultaten van modellen, van ketens van modellen, dus niet alleen maar meetgegevens.	rd herkomst

10:7	Interviewer: En over wat voor orde grote praten we? Interviewer: Wisselt. De meeste modellen nu zijn orde grote van een paar gigabyte aan inputdata voor die modellen. Output data, nou 1 van mijn afstudeerders, die had 60, 70 gigabyte? Dat is ongeveer het grootste wat we tot nu toe gehad hebben. Wat toch een drama is, want dat is niet goed te doen met Git en dan moet je toch 1 of andere bigstore gebruiken. Interviewer: Gesodemieter. Dat is ongeveer de grootste die ik tot nu toe gehad heb voor één publicatie, een goeie vijftig gieg, gezipt.	opsl capaciteit opsl code
10:8	Interviewer: Waar sla je dat dan op? Geïnterviewde: Zij heeft altijd bij IBM gewerkt. Dat is wel handig als afstudeerder iemand die 6, 7 jaar gewoon als programmeur in Amerika bij IBM gewerkt had, en daardoor nog een professioneel Dropbox account had. Ze had haar data daardoor gewoon volledig.... Ze had haar data volledig op haar laptop staan, met een backup plan lokaal thuis En ze heeft in die tijd met mij, de hele dataset gedeeld via Dropbox en ik heb het lokaal staan ik heb het nergens anders staan. Interviewer: Is dat ook iets wat je zelf gebruikt Dropbox? Of was het nu toevallig een keer? Geïnterviewde: Ik gebruikt Dropbox voor papers, niet voor data. Ik gebruik vaak bigstore functionaliteiten in Git. Dat komt als de datafiles 100, 150, 250, 300GB per datafile zijn. Dan is bigstore functionaliteit binnen Git prima en het voordeel daarvan is dat ik gewoon 1 repository heb waar ik alles heb. De pest is zodra ik een aparte oplossing nodig heb voor mijn data en een aparte oplossing voor mijn code, is de barrière van het kunnen delen met anderen… die neemt toe, dan moet je altijd 2 dingen syncen en dat is pestilent. Dus dat is een probleem het is altijd een beetje zoeken naar oplossingen.	opsl capaciteit opsl code opsl ext datahouder opsl share and sync
10:9	Interviewer: En voor delen? Deel je zowel je data als je code? Geïnterviewde: Dat hangt er van af, bij publicaties is het zo dat ik eigenlijk altijd probeer een aparte github repository te hebben. Dat ik een aparte repository in Github kan hebben, gekoppeld aan een paper en daar staat dan alle code in. Ik probeer er i.i.g. voor te zorgen dat de codes die gebruikt zijn om alle figuren in een manuscript te preproduceren er ook in is verwerkt, zodat de data die nodig is voor het reproduceren van de figuur in de github repository zit.	del code opsl code rd code rd levels
10:10	Geïnterviewde: Dat het kan zijn dat die data het resultaat is van een aantal processing steps op een heel grote dataset, kan betekenen dat we die grote dataset niet plaatsen op GitHub. En dat heeft dan weer te maken met die big store. Het andere probleem heeft er vaak mee te maken, dat ik die data gegenereerd heb met simulatiemodellen waarvan ik de onderliggende data ook niet altijd kan delen. Dus dan is het vaak dat je zegt: Ok, Ik gebruik dit simulatie	del code del motivatie opsl code

	model zie dit paper en daar moet ik stoppen. Het is niet aan mij om dit model te delen met anderen.	
10:12	Geïnterviewde: Het gaat wel interessant worden he, want de nature journals gaan het meer en meer eisen. Ik ben heel nieuwsgierig hoe dat zich gaat ontwikkelen. Ik heb het zelf wel al een paar keer gedaan dus, dat ik zei: het is allemaal leuk en aardig, maar geef me nu maar even het model.	ind werkwijze
10:13	Interviewer: Maak je ook zelf metadata aan gebruik je op 1 of andere manier metadata? Geïnterviewde: Dat zou ik moeten doen. Nee ik heb vaak aan een readme achtig iets met interpretaties. Maar ik heb vaak een data folder met een readme erin wat er in die data folder zit en ik probeer in de naamgeving van de files, dat zijn vaak tarballs, vaak zo discriptief mogelijk te zijn in wat het is.	dd readme vlg ja
10:14	Geïnterviewde: En heb je een eigen naamgeving? Of is dat iets wat jullie met de afdeling gebruiken? Geïnterviewde: Ik heb mijn eigen conventie maar dat is dan bijvoorbeeld ok: n experiments x strategies zoiets weet je wel, dat soort dingen Interviewer: En je vraag niet aan je phd's en je master studenten om dat ook te doen? Geïnterviewde: Ik geef ze dat als best practice mee. In die zin wel.	nmc persoonlijk vlg naamconventie
10:15	Geïnterviewde: Maar aan die kant is denk ik wel het een en ander te winnen. Het is een discussie die in de wetenschappelijke community waarin ik actief ben speelt, hoe we daarmee om willen gaan en men is nog zoekende hoe we dat willen doen. Het is wel zo dat als ik iets van een andere onderzoeksgroep moet reviewen, dat ik binnen enkele minuten snap wat er gebeurt. Dat is goed genoeg.	nmc persoonlijk
10:16	Geïnterviewde: Ja daar gaat die provenance echt een rol spelen. Interviewer: En hoe pak je dat dan aan? Hoe zorg je dat het volgbaar blijft? Interviewer: Wat ik dus probeer te doen is gewoon Notebooks, dus met Jupyter notebooks werk ik veel. Daar probeer ik dan heel netjes die provenance in vast te leggen. Ik pak mijn dataset met de experimenten, ik doe deze, deze en deze processing steps en dan schrijf ik het weg naar een nieuwe date file met een andere naamgeving. En dan kan iedereen dus zien, ok, Dit zijn stappen die uitgevoerd zijn. Dat is hoe ik het probeer te doen.	dd data bewerken dd eln dd inhoud vlg ja vlg transformatie

10:17	Interviewer: En dat is ook versioning wat je dan probeert te doen. Geïnterviewde: Ja daarmee doe je ook versioning. Interviewer: Is dat dan versioning die je uit github haalt? Of is dat in dit geval gewoon dat jij v2 achter een bestandsnaam zet? Geïnterviewde: Nee in principe, meestal is het zo dat, je doet een aantal iteraties en dan ga je echt van: dit is de definitieve set van experimenten die ik ga doen en die run je één keer omdat, dat kan rustig drie weken duren of zo. Dus die run je één keer en dan kun je het verder met naamgeving oplossen. Het is niet zo dat ik er v2, v3 o.i.d. achter zet, daar word ik helemaal gek van. Gewoon netjes bijhouden in Git met goeie heldere commits dat snap ik. Dan kom ik er wel uit. Maar je moet niet V1, 2, 3, 4, 5, 6, 7, 8 dan haak ik af. Interviewer: Je gebruikt die van Git dus? Geïnterviewde: Ja	nmc persoonlijk vlg code vlg ja vlg naamconventie vlg transformatie vlg versie
10:18	Interviewer: Folder structuren? Gebruik je die ook? Geïnterviewde: Ja elk project heeft bij mij een identieke folderstructuur en al mijn afstudeerders en promovendi moeten die gebruiken. Interviewer: Dus die leg je wel op? De naam conventie niet maar de folderstructuur wel? Geïnterviewde: ja ik zeg: oké je hebt main, de root dan heb je een data folder Je hebt een figurenfolder en je hebt een code folder en dan is het gewoon klaar.	fdst groep vind folderstructuur
10:19	Interviewer: Nog heel even terug naar die dataset en het labnotebook. Maak je ook links in je labnotebook naar zo'n dataset? Jupyter gebruik je he, zei je? Geïnterviewde: Ja, nou in mijn code staat gewoon een note en daar staat 1 of ander relatief pad. Interviewer: Je zou vanuit je notebook rechtstreeks naar een dataset kunnen klikken? Geïnterviewde: Niet echt kunnen klikken maar je leest hem in.	dd eln dd link vind ja vind notebook
10:20	Interviewer: Waar de data staan hebben we het een beetje over gehad dus die.... eh veel staat op github, geldt dat trouwens ook voor de input data sets die je aangeleverd krijgt? Zet je die ook op Git?	opsl code
10:21	Geïnterviewde: Dat hangt er van af hoe groot ze zijn. En in hoeverre ik het model op github kan zetten, soms kan dat helemaal niet en soms mag dat niet. Dan gebruik ik vaak en TU Delft SVN of een TU Delft Git, volledig private, en doe ik het op die manier maar dan ben je weer terug bij het punt van de hele grote datafiles die je dan moet gaan rondpompen. Mijn promovendi hebben verschillende manieren om daarmee om te gaan, hangt ook een beetje van hun case af bijvoorbeeld. Een van mijn projecten was een samenwerking tussen Civiel, werktuigbouw en hier en daar had ik gewoon een SVN achter hangen van de TU en daar hebben we alles mee gedaan.	opsl code

10:22	Geïnterviewde: En we hebben ook wel een tijdje gebruik gemaakt van de large file store van de TU. De bulk. Op een gegeven moment hadden we dingen op bulk staan	opsl centraal
10:24	Geïnterviewde: Tegelijkertijd het feit dat je twee repositories moet syncen is een extra barrier in hoe je dingen doet en dat is gewoon irritant. Hoewel het ook onvermijdelijk is for the time being.	indmot mate van overhead
10:25	Interviewer: Dataveiligheid, hoe zorg je ervoor dat er alleen de mensen die er bij moeten kunnen bij jouw data kunnen? Geïnterviewde: Niet. In die zin dat als ik werk met proprietary modellen en data. Dan zal ik gebruik maken van een lokale TU Delft gerelateerde data oplossing zoals een bulk of van dat soort dingen. Maar ik heb niet te maken met medische data of persoonsgebonden data in de regel. Dus het is allemaal minder gevoelig wat dat betreft.	bvlg autorisatie
10:27	Interviewer: En is het voor jou interessant om te zien wie er aan de data gezeten heeft zeg maar wie hem gekopieerd heeft of er in gekeken heeft? Zou je daar logging van willen zien? Geïnterviewde: Ik denk dat hoe we nu met onze data omgaan is een beetje nonchalant. Anderzijds is het een beetje gebaseerd op wederzijds vertrouwen in de afdeling en dat soort dingen hier. Binnen mijn sectie zit er… binnen de afdeling zijn niet zo veel mensen die zomaar even op de bulk kunnen.	bvlg autorisatie logging vlg share
10:29	Geïnterviewde: Dat hebben we eerder gedaan en is niet specifiek voor dit project. In een eerder project is gedeeld met de VU en daar werd volgens mij gewoon gedropboxed. Of misschien Surfdrive ook wel.	del sync and share
10:30	Geïnterviewde: nou ja, 1 linkje uit Surfdrive en toen hadden ze het en dat werkte wel lekker.	del sync and share
10:31	Interviewer: Encryptie, heb je dat wel eens nodig? Gebruik je het wel eens? Geïnterviewde: Ik heb het zelf nooit gebruikt. Dat wil niet zeggen dat ik het niet zou moeten gebruiken. Bijvoorbeeld GPDR dingen met bijvoorbeeld cijferlijsten, die zouden eigenlijk encrypted op je harde schijf moeten staan. Interviewer: Je mag nu alleen toch maar een cijfer aan een studentennummer toekennen? Geïnterviewde: Maar ik heb natuurlijk veel meer data dan dat. Formeel zou ik dat dus moeten encrypten. Maar ik heb een Mac, dus het is volledig encrypted [lacht]. Interviewer: Maar dat is het natuurlijk wel als je harde schijf gewoon encrypted is. Geïnterviewde: Volgens mij draai ik hem in encrypted mode, ik dacht het wel.	bvlg encryptie

11:1	Geïnterviewde: Well, the question is what what do you consider data right. I mean we have this software that contains essentially models, which you could also consider is data, because it's a concept of how the world works, but it also contains like parameters, which are data. And are important too. The parameters basically define essentially the models and the modelparameters define what the result would be.	rd definitie rd herkomst
11:2	Geïnterviewde: then I would consider software part of the data somehow. Because the software itself contains important information about what we actually did.	rd code rd definitie
11:3	Geïnterviewde: So then we produce all that. I mean the simulation produces internally huge petabytes of.. or even more amounts of data, be stored and sometimes infrequent snapshots.	opsl capaciteit rd simulatie
11:4	Geïnterviewde: So we'd be store snapshots, which can be of the whatever flow field, that is velocity, pressure or whatever concentrations, these kind of things. Geïnterviewde: In whatever intervals which basically depend on what what you want to do. If you want to do a visualization then you would like to have every second a picture or so. Interviewer: So the data you get is data and you create by simulation? Geïnterviewde: By simulation yes.	rd levels
11:5	Interviewer: So the scripts to create the data are part of the data as well. Geïnterviewde: I would say so yes, because the data only makes sense, if you know how it was created. So the interpretation, otherwise it's just numbers.	dd inhoud rd code rd definitie
11:6	Geïnterviewde: Well I mean it's a software which involves implementations of models. So what we actually store in terms of data, is like this as I said the snapshots of instantaneous solutions, but also often only just statistics like whatever time averages, or spatial averages or whatever spacial domain.	rd definitie
11:7	Geïnterviewde: Instantaneous snapshots can be a few megabytes, a few terabyte. Can be a lot, I know a PHD student who produces like 40 terabytes snapshot.	opsl capaciteit
11:8	Geïnterviewde: Well you could could produce several of these snapshots per per minute, but you cannot store them, so you don't do that usually and it's basically challenging to handle this. But mostly we are only interested in the final product and usually that is either a visualization of the instantanious thing which s nice to provide, but the hard scientific facts are just statistics. Which is then a relatively small amount that you can essentially present in one figure or table.	opsl capaciteit rd levels

11:9	Geïnterviewde: Then of course there's the experimental people which I don't feel directly part of, but they do well a lot of measurements, where there's a lot of hardware post-processing before actually something happens. But then there's also a lot of software put post-processing afterwards. And they also do a lot of optical techniques where they first record an image and then do some algorithms on that and then extract some velocity fields or pressure fields. So it's very diverse and I mean also in terms of what it actually represents and what's the size and the frequency that we need that.	rd levels
11:10	Geïnterviewde: Yeah I mean...The software is managed in a Git Repository so that we can go back and forth and have everything documented. So I mean it's also the sharing part.	del code opsl code vlg code vlg ja vlg transformatie vlg versie
11:11	Interviewer: So if you run a simulation twice, you get the exact same data? Geïnterviewde: Exactly the same result.	rd simulatie
11:12	Geïnterviewde: I mean if you have the software itself and the definition input files. So we try to store both: Software is always in Git and then the definition files. That's something less well defined. I try to keep it all and collected but sometimes things get lost.	opsl code rd code rd definitie
11:13	Geïnterviewde: We stored like a really large database that we use for another simulation s an inflow condition and then see what happens downstream. And this isn't something which I cannot store in my laptop anymore. While usually the input files yes and the git repository.... But this inflowdata is then sometimes difficult to find at least. I mean it should be in on some hard disk somewhere.	opsl code opsl ext datahouder
11:14	Interviewer: And do you put it in Git as well for the versioning part of it. Geïnterviewde: I mean the software, yes. Binary files I don't put in Git, because it's just crazy because then the epository explodes.	vlg code vlg ja vlg transformatie vlg versie
11:15	Interviewer: Where do you put the binary files? Geïnterviewde: Well there are the files in Munich which Icannot access anymore because that's when you change universities you lose your account. Theoretically it would be findable, accessible and whatever, but it's not in practice because the administrators say well you're not working here anymore, you don't have an account. There's a lot of data in Amsterdam right. I don't know where physically exactly but we know how to access it.	vind ja

11:16	Geïnterviewde: Well usually I mean when you do big simulations, the majority of the data stays on the computer where we compute it. And then we transfer only partially post-processed things or, if there's a problem usually then also larger amounts. But we try to avoid this, because it takes a long time. Then you have the problem where do you store?	archief waar archief wat del datatransfer opsl ext datahouder rd levels
11:18	Geïnterviewde: sometimes when we do paper writing I put it in the Git repository for day to data text processing and. So for many of my papers it's basically part of the publication. I sometimes also publish but that would be on my personal drive.	ind werkwijze opsl code rd definitie vlg versie
11:20	Interviewer: Oh OK but not on TU drives? Something like bulk for instance? Geïnterviewde: Not not so much. I find it not so convenient to use with Mac. I know I mean the Windows people they always talk about the K drive, but for me I have to mount it every time and then it somehow gets inconvenient.	indmot mate van overhead
11:21	Interviewer: What kind of cloud drive do you use? Geïnterviewde: Still using Dropbox. Interviewer: But that's small. You have two Gb right? Geïnterviewde: I have a terabyte. Interviewer: Ah you bought extra?	ind werkwijze opsl share and sync
11:23	Interviewer: And you said dataset stay mostly on the system itself. So what about reliability if the system crashes you lose your data? Geïnterviewde: Well. That's possible, that sometimes happens.	opsl betrouwbaarheid
11:24	Interviewer: But it's not really bad because you can easily reproduce it. Geïnterviewde: Sometimes not. Not easily. Right. I mean when you have spent like whatever 20 million core hours for computing or something.	bvlg backup opsl betrouwbaarheid
11:26	Interviewer: And what about metadata? So that's something that's important for you for your datasets? Geïnterviewde: Welldata isstored usually in a binary format so you cannot read it. You need some instruction to know how to read it first. Interviewer: But is that metadata or is it more like data documentation? I mean enter data in my mind right. In my mind metadata is more like a who created is, when was it created, for what projects? Geïnterviewde: I think that if we have something really important that is useful for others and yeah we just format it in ASCII format or something else, or provide a small Fortran code or something to a team that reads the file so they can see how to use it. And then of course it gets a header like please cite this publication. So yeah. I mean we don't do this like everyday because well it is too much.	dd inhoud md definitie md ja
11:27	Interviewer: Uh a little off topic but where do you archive your dataset. Would you put it at Zenodo for instance Zenodo from CERN or 4TU? Or you don't. Geïnterviewde: Well you usually as I said important for the data we usually use just te archive system like the	archief ja archief waar

	LSE in Munich or SurfSara and now we also did something with the 4TU server and then,let's see. I think usually it's just on my Mac	
11:28	Geïnterviewde: I mean sharing is something, I mean a lot of websites also where you can just publish data. I mean you can attach it to a journal or 4TU or upload it onresearchgate. Interviewer: But that is the enddata. Geïnterviewde: This is the end and what I do is something. I usually to share put in on some disk inthe network where everyone can access it.	del centraal del intern
11:29	Interviewer: Then you're shared with the people in this building? Geïnterviewde: Depends. I mean generally everyone with access to HPC12 could get access. Interviewer: Is that only people from TU or isthat people from outside as well? Geïnterviewde: That should be almost almost exclusively people from from within.	del intern
11:30	Interviewer: So there is no need for you to share your data with somebody in, I don't know, the US? Geïnterviewde: Sometimes there is, and that's always a problem. And then you put it on some FTP and then you try to transfer it but it's always an ad-hoc solution.	del centraal del datatransfer del extern
11:31	Interviewer: I suppose if this data would be on your Dropbox you would just use that? Geïnterviewde: Yes then it would be easy to share it.	archit share and sync del sync and share
11:33	Geïnterviewde: Source Code is a different thing and thus I like to. I mean I am also open to share. But I like to keep track of who has what.	logging
11:34	Oh so could it happen that you share data with somebody else and you want him or her or the institution to write some sort of agreement. Geïnterviewde: Myself not but many collaborators, especially companies, but also Dutch research institutions like would have a Marine and NWO, they are crazy about their non-disclosure thing and they don't like to share anything apparently.	bvlg contract
11:35	Geïnterviewde: So I'm more open. What I generally... well I put it somewhere on a website and I know that it's for everyone and I don't care. So I want that as many as possible have that. Sometimes you could also post a code on github or so that it is shared. But generally, most things I give personally to people, virtually personally, and that's based on trust that they would ask me if they can give it to someone else.	del code ind werkwijze logging

11:36	Interviewer: And do you keep track of how many people downloaded for instance your data. Geïnterviewde: I mean my main code they're part of the license agreement is that everyone should just use his personal account, so I know how many people have a personal account. I know that it's violated of course ,that people give passwords to someone else or. Yeah it's not I don't think that this is nice. I mean it's. I mean I'm open, but I think trust is important. But for whatever github or researchgate, I don't know if it's possible to track that.	logging vlg share
11:37	Interviewer: It's not something that's important for you? If it would not be there.... Geïnterviewde: It's not important, sometimes it's interesting to see that one paper is downloaded a lot. But that's not so important. What's more important is when they cite it, or use it, what they say about it.	del motivatie logging
11:38	Geïnterviewde: Well my data is basically....Input directory where input data is and an output directory where output data is. It's not allowed to modify the input data. So that makes kind of sure, that I always know what was the input and what is the output. Of course I mean when you talk about something like the hard disk fails and now the data is somehow corrupted. Unless I get an error message I don't have much means to knowledge to find this out.	vind folderstructuur vind ja vlg folderstructuur vlg ja
11:40	Interviewer: So in your case you just have it sorted in input and output? Geïnterviewde: Yes and then in terms of statistics well the file name has a step number in the file name,also has a time stamp of course, or the timestamp of the operating system, but it could of course change if you copy it, but the filename itself compute puts the time step of the actual simulation.	nmc persoonlijk vind folderstructuur vind ja vlg folderstructuur vlg ja vlg naamconventie
11:41	Geïnterviewde: during the active research base the metadata would be the definition file of the of the software that generated the data.	md definitie md ja
11:42	Geïnterviewde: And so I mean I run my code and basically read some input file, but this input file only contains things that you want to change. And there is an outputfolder that includes the Git commit number. And then all the possible settings that you could have changed. But then the default. So that this file could be also used as an input file again to exactly reproduce that so that. So that's basically all a meta data.	md definitie md ja vlg folderstructuur vlg ja vlg transformatie
11:43	Geïnterviewde: I would like to have something because when a master student stops. Then I get like this hard disk which I then put somewhere. Interviewer: You get a hard disk? Yeah. Oh of course he has big datasets. Geïnterviewde: And then yeah. Then I put it somewhere. I mean I have a hard disks from all my professional life. Probably some I cannot mount anymore because the one that they're formatted with whatever old Linux file system formatted.	archief eisen archief ja archief waar archief wat archit ontwerp ind werkwijze

	And. I don't have a workstation where you could put it in but maybe even the file system form it doesn't doesn't work. Yeah I mean at homeI have a file server where I can store things but this is just just too much I don'tput it there. It would be nice to have something central where things could be,...where we could just copy things. And then know that it's somehow back-upped. So that we. Can delete the copy that we have ourselves we can rely on that it's there. I mean I copied it there I have a whatever directory name like Master thesis from Ali, and I can find everything Interviewer: And Ali comes from Iran of course.. Geïnterviewde: No Ali is actually Danish. I have the hard disk from Denmark, anyway but that would be nice because I mean some of my colleagues they also make a copy off that hard disk from a student and take one copy home to be sure... you shouldn't have 'a copy on site' Interviewer: But this more like archiving right. Yeah. You will never do anything again with the data? Geïnterviewde: If you use it never again. And it's not important, no one would ever want to see that record. then you can delete it. OK. Interviewer: So how long would you keep it then? Geïnterviewde: I would say usually at least ten years	
11:44	Interviewer: OK so it sounds pretty simple. In your case you have this input file and outpu file, the data you stored is locally on a machine that does the computations and well basically it stays there forever. You maybe take it off when it's full or something the disks? Geïnterviewde: Well I mean then I mean if somebody..... I would just put it to a tape archive and then.. Interviewer: OK and then for 10 years? Geïnterviewde: I think so yeah.	archief eisen archief ja archief waar archief wat
11:45	Geïnterviewde: Yeah I mean everything which is produced with one software has a name convention but. No I know colleagues which which do that. When they have,... also when they write papers everything has to be exactly the same, but I like to keep a bit of individualism that the PHD students also have and also would not be great if I tell them how to name their files.	ind werkwijze indmot mate van flexibiliteit nmc persoonlijk
11:47	Geïnterviewde: And what I always tell people is when they do many simulations they should have something like your Excel table where they put in just a list of their simulations. With parameters that they modified and where it is and when it started, when it finished and things like this. And some some also basically put in there some result picture of just something. But I don't have a systematic approach for this and I never, I mean I always have my notebooks but I never really did it systematically.	dd data verzamelen dd eln dd inhoud ind werkwijze vind ja vlg ja

11:48	Geïnterviewde: Yeah one thing, for the practical work is always like, how long does it take to access data. Right. How long does it take to access it. So what's basically the latency time. For example one of my PHD students who is not finishing he's making nice visualizations. It's very nice and you can also make a movie so he's downloading like a few time, instantaneous snapshots from tape on the project drive, we have a 100 terabyte projectdrive until it's full. Then he goes to the visualisation station where there are pictures produced, and then there are some pictures produced. The the project drive is deleted, new datasets are pulled from archive whatever three more pictures are generated and that's a very slow process, of course you can can script it somehow. It's a very slow thing. And so that's when you have the big data and of course my hard disk system where too many people have. That's that's the latency is also very .... [hard disks fromstudents he keeps in a cupboard] I have to find it first and then have to connect it and then look at it and ..... No I mean there's something something like medium there's a medium capacity which is really on the network. I never really fully explored the the two Delft infrastructure. You have for some groups. We have shared drives there. They seem to work well but I'm not really sure what's the capacity, if I could copy there.	archief eisen archief waar archief wat archit ontwerp del datatransfer ind werkwijze
11:49	Geïnterviewde: Yes so I mean we still do a lot of simulations on the HPC 12 cluster and then you know the filesystem is slow it's also not very big and it's permanently full. Also I mean active data is things for me before it has been published in The Journal. So during that time we really need somehow really quickly accessible, it's basically sometimes daily accessed	archit ontwerp del datatransfer
11:50	Geïnterviewde: And I mean for Surf Sara for a small pilot project where I just write an email and say I would like to test the system and if I write 8 terabyte, that's basically the default thing. And if I want to have more then I should justify it. So that's that's a really small amount	archit ontwerp
12:1	Geïnterviewde: Yeah. So I know from experience of trying to help people with electronic lab notebooks. In a previous job the computational scientists have a completely different way of recording their progress. Those of us in the lab we tend to... there's still a lot of people with their paper route. Geïnterviewde: So yes there's quite a different way of looking at recording and not recording what you're doing and get top in all of these other computers in control.	dd eln dd papier
12:2	Geïnterviewde: Temperature and pressure and and the flow rate. But this is where the lab notebook becomes important because you need to keep a note of exactly which fluid rate you're setting. And then you measure the pressure coming out.	dd data verzamelen dd inhoud ind werkwijze rd eln

12:3	Geïnterviewde: This is the computer so we control everything our flow rates and all the pumps are controlled by a computer. So I can have the data recording on here. Well I can have the pump control on here I record my pressures on here and then I can access the electronic lab notebook.	ind werkwijze
12:4	Geïnterviewde: And what's nice is that you can enter some meta data there. And what we tend to do it's become a kind of within the research group. We tend to automatically put a fairly descriptive file name on it. It's the kind of thing that you think one person starts doing and it's kind of spread. So automatic so you can set up a kind of template for the file name and then you just change your put in the individual. I can show it to you if you want to later.	ind werkwijze md ja md vorm nmc groep vind naamconventie vlg folderstructuur vlg ja vlg naamconventie vlg transformatie
12:5	Geïnterviewde: No no no no. So what is within the data recording those if it's just where this is where I hope I get my terminology right. We can enter some meta data which I only discovered how to do this about six months ago where you can you there's that you can set up a little form so that every time you run a test it says okay what's the temperature. What's that rock type. What's your flow rate. So you can put some of the things that you would normally written down. You can make sure that it's actually attached in the file with it.	md ja md vorm vlg ja
12:6	Geïnterviewde: And to me it my understanding is metadata is the data which is what describes the data. So. So what we do. So when we can run a test. So this is gonna be recording the time and the pressure and the temperature but then it's got a little bit in the header then where it'll tell me what what was the oven temperature which is fixed all the way along. So that kind of the things which are the sort of the basic properties which are fixed then that's in a little light comes in a header. So if someone comes back and looks at and I've actually found this really useful	md vorm vlg ja vlg naamconventie
12:7	Interviewer: And do you use something like data documentation like really some sort of text document describing your data. So other ones can understand what it is or is it so obvious that everybody ..... Geïnterviewde: No, I think that would be assumed that would be what your Labbook would do is. So we don't have any we don't have any kind of formal documentation or	dd eln dd inhoud
12:8	Geïnterviewde: We don't tend to share data in that way because it partly because I don't know maybe it's historically because the experiments we've got a lot of industrially funded.	bvlg contract del motivatie

12:9	Geïnterviewde: my data is for [company name] and I wouldn't share my data with somebody from [other companies]. Geïnterviewde: So there is that kind of rule, a certain area of confidentiality.	del motivatie
12:10	Geïnterviewde: So every time you move you're like, if I move is my data going to die as I leave so you become much more interested in trying to. So I tend to be quite obsessive in labelling my directories. So if somebody comes after me they should hopefully be able to go. OK. Right. She was doing the steam foam tests and then here's the data. And so hopefully they should be able to follow my life just to be over the label directories, but this is why I got interested in the electronic lab notebooks because how did you leave your lab notebook behind	dd eln
12:11	Geïnterviewde: but it's always difficult to because especially paper notebook if people don't know where the information is you know unless you've got some kind of indexing system it's almost impossible to find information in a paper one and this is why I'm quite enthusiastic about the electronic lab notebook because for me it solves some of these problems because, it's searchable so it should be a lot easier for people to find information and they should be able to come through.	dd eln vind ja vind notebook
12:12	Geïnterviewde: but at least people can only least I can search at least somebody's coming after can search. And it's a bit it's a lot more accessible than here's. An A4 notebook which is filled with my handwriting.	vind ja vind notebook
12:13	Geïnterviewde: we would share the data in terms of publishing the factual output but not the the raw data. Possibly if I'm writing then I'll share I might share the raw data with the colleagues thatI'm writing a paper with.	del motivatie rd levels
12:14	Geïnterviewde: So where the data is all on the computer that runs experiments so all the kind of raw data. Then everything else is on my computer.	opsl ext datahouder
12:15	Geïnterviewde: I keep it all in my documents which is the central folder? Geïnterviewde: This is where this is where I'm a friend of Cook from doing something really stupid. Interviewer: I'm not sure. I think my documents is.... you have windows 7 still? Geïnterviewde: yeah. Interviewer: Then it's linked to your home folder. Geïnterviewde: Yes.	ind werkwijze
12:16	Geïnterviewde: I've either had access to a centralized folder where I could back it up, because I know the importance of backing up data.	bvlg backup

12:17	Geïnterviewde: I use dropbox for smaller personal files or occasionally if I'm working on a paper for something that's Yeah. I've never thought, for me Dropbox is because it's external. And also Yeah I again because of a lot of the funding for these projects as it was either oil companies or when I was in Paris I was funded by [name]	del sync and share
12:18	Geïnterviewde: I know that I need to back up my data but the idea of keeping it essentially in ELN is I love being able to... so I can be at my experiment I can input the data here.	bvlg backup
12:19	Geïnterviewde: Quite often because my laptop lives in the office and I usually email the data straight up so I can finish the experiment I go upstairs I can start processing the data and then I can access the ELN	ind werkwijze opsl ext datahouder
12:20	Interviewer: You were talking about backups earlier. Did you.... Have you ever.... Did you ever have to do a restore? Geïnterviewde: No I've been very lucky. Okay. Um I think this is going back to the old days of the old three and a half inch floppy disks. you know in your piece do you know your documents were on the old we still got some of the computers here we've still got the old ones.	bvlg backup
12:21	Geïnterviewde: Yeah I know how important data is and losing the last thing you know you can if you if you lost your data and you sort of lost three and a half years of work or a year's worth of research data it would just be a disaster.	bvlg backup
12:22	Geïnterviewde: Well let's sort of sync you know up my computer once a week put it to the external hard drive and but there's all the different options and which is the best option? Which one should I be using?	bvlg backup opsl ext datahouder
12:23	Interviewer: What about data security. You need to keep it secure from other people for instance? Interviewer: I think yeah. Geïnterviewde: Not like really strictly secure but I think so we I think in general we assume that if the data is on a lab computer only people are authorized to be in the lab could be in here technically. Geïnterviewde: Could somebody who was working for another oil company come in and just like trying to secretly steal the data. But if they didn't know what was going on in the experiment, so raw pressure data is just like a noisy you know it could be just a noisy signal unless you know exactly what is going on in that experiment. So you kind of you need the data and you need the notebook at the same time.	bvlg authenticatie
12:24	Geïnterviewde: I think it's more of the processed data which is like with the knowledge of what the settings were in the experiment and what was going on and all of the parameters that were being varied the process data is quite important. Geïnterviewde: So we do tend to be like I wouldn't share process data with everybody.	del motivatie

12:26	Interviewer: but it's your laptop for instance encrypted. Geïnterviewde: No. Interviewer: Do you use encryption for any patenting. Geïnterviewde: No I don't.	bvlg encryptie
12:28	Geïnterviewde: But I am aware when I'm working on a project you know you don't just share your data. Near process data with anybody. The process data is just for your and your industrial partners. If you're writing a paper quite often you would anonymize stuff.	bvlg autorisatie del motivatie
12:33	Geïnterviewde: You can prepare a graph and or you keep some a table of data as you're going along but the data was held on a local very secure server and then the electronic lab notebook was literally just a replacement for the paper	rd eln
12:34	Interviewer: You have your own naming conventions and folder structure. Geïnterviewde: Yes. Interviewer: That's not departmental.....? Geïnterviewde: It's my I my personal one	ind werkwijze nmc persoonlijk
12:35	Geïnterviewde: So I use a dating structure. I started that when I was using a lab notebook a paper notebook because I actually started putting tabs on each week with the day of the week so I could usually flip can go okay. It was in the summer and then I could. So I was using that as a linking the data as a linking mechanism on the experiments. I see. So we can there's a few different things that you can vary so I'm looking at high temperature. I'm working with different surfactant so I'll quite often I'll put the surfactant in the temperature.	ind werkwijze vind ja vind naamconventie vind notebook vlg folderstructuur vlg ja vlg naamconventie
12:36	Geïnterviewde: And then it can auto fill in the data for you will get. Which again is very useful because then you can link to your lab notebook. What was the day I did the test and then you can search. So it's all attends to auto fill in the date which is really useful.	archit eln dd eln dd link
12:39	Geïnterviewde: On the most basic on the most basic version because this throws out a text file with raw pressure and then I want the pressure differences so I I just open up an Excel to a couple of basic calculations and then see the I's calc so I have a count. So I have the raw data and the calculations files so that's my versioning. It's possibly the most basic.	vlg ja vlg transformatie
12:41	Interviewer: I can imagine that you have some sort of experiment the first four go wrong and a fifth goes right. So what do you do with the data sets of the first four. Geïnterviewde: They're not actually worth anything so I'll I'll have them and I'll keep the raw data set and I'll know in my notebook I'll know that the these experiments on these dates were the ones which went horribly wrong.	dd eln dd inhoud

14:2	Geïnterviewde: Onze data zijn de codes die we ontwikkelen, zelf, de computer codes, en natuurlijk wat daar in en uitgaat. Zou ik denken ja, dat kan een mesh zijn dat kan data van een patiënt zijn. Dat kan van alles zijn	rd code rd definitie rd herkomst
14:3	Geïnterviewde: Dat kan zoiets simpels zijn als text files als input Parameters Settings. Dat zijn veel soorten files.	rd definitie
14:4	Geïnterviewde: Maar ook wel op het medische vlak: CT scans..... Interviewer: En dat zijn echte CT scans? Geïnterviewde: Jaja dat zijn echte, of we maken ze zelf, model CTs. En in het nucleaire vlak zouden dat misschien meer geometrien zijn, een mesh, gemeshte geometrien.	rd herkomst rd meetdata
14:5	Interviewer: En als je het hebt over parameters, dan zijn dat parameters die je in een model stopt, waar een dataset uitkomt? Geïnterviewde: Ja, dat zegt hoe fijn moet het zijn of hoe moet het gediscretiseerd worden. Parameters, over het algemeen zijn dat soort inputfiletjes klein. Een hele simpele... dus als je je data wat er uitkomt nou kwijt zou zijn, dan zou je dat kunnen hergenereren door je input nog eens een keer door de code heen te trekken. Ben misschien wel een week verder. Maar goed dan heb je het wel.	rd herkomst rd simulatie
14:6	Interviewer: Dus jouw ruwe data kunnen dat soort inputfiletjes zijn? Geïnterviewde: Ja,ik kan de andere data weer genereren. Code niet en input ook niet.	rd levels
14:7	Geïnterviewde: CT natuurlijk ook, die verzinnen we ook niet zelf, meestal komt die ergens vandaan. Komt meestal van samenwerking met Erasmus of hier met de kliniek.	rd herkomst rd meetdata
14:8	Interviewer: Waar moet ik aan denken bij code? Geïnterviewde: Je krijgt gewoon een computer code. Fortran, C dat soort dingen. Interviewer: En daarin programmeer je hetgeen je gebruikt om die parameters om te zetten? Geïnterviewde: Dat zijn de fysische modellen. Je moet meestal denken aan grote eindige elementen achtige codes	rd code
14:9	Interviewer: Dus we hebben het al een beetje gehad. Wat is er relevant om op te slaan. Dat zijn dan de input files en de codes? Geïnterviewde: Nou dat is het allerbelangrijkste. Maar kijk. Om nou dat alleen maar op te slaan is niet verstandig, maar we slaan ook wel....eh... bijvoorbeeld als je transientanalyse doet, begin je met de steady state. Het is wel verstandig om zo'n steady state op te slaan want anders blijf je....eh tenminste tijdens het project van een paar jaar dan slaan we dat sowieso op. Zo'n volledige transient is sowieso niet op te slaan, dus alleen maar snapshots om de zoveel tijd is een shot van hoe het veld eruitziet, het is gewoon veel te groot om allemaal een tijdstaf op te slaan. Dus daar maken we sowieso al een keuze.	rd levels

14:10	Interviewer: En moet je ook in staat zijn om back ups en restores te maken? Of is het zo van je maakt voldoende stappen in zo'n onderzoek dat als je een keer een stap verliest dat dat niet erg is? Geïnterviewde: Nou kijk, het gebeurt natuurlijk niet vaak. Dan zouden we wel weer kunnen hergenereren. Deze spullen staan allemaal op de hpc's daar wordt volgens mij een back up van gedraaid want de meeste stukken tenminste op sommige stukken. Geïnterviewde: Maar ik hamer er bij iedereen op dat ze hun eigen spullen netjes weer ergens moeten zetten want.....eh ja	bvlg backup opsl ext datahouder
14:11	Interviewer: Waar staan jullie data allemaal behalve op de hpc? Geïnterviewde: Denk het meest op de hpc's. Geïnterviewde: Dan heb je nog de individuele storage van de AIO's en hun eigen h schijf. We hebben ook Project schijven, dat is nog niet zo heel oud, sinds anderhalf jaar, dus dat doen we nog niet zo lang. Dat gaan we wel steeds meer doen voor een project een schijf aanvragen. Dat is het eigenlijk wel. Buiten dat hebben we natuurlijk ook nog de dropboxen die worden veelvuldig gebruikt. Ook zeker met uitwisseling tussen mensen. Sommige van die projecten zijn Europees en er wordt wel eens wat uitgewisseld, of via de schijf van de TU of via dropbox of mail. Maar meestal zijn dat natuurlijk niet de data zelf. Maar er worden gewoon andere dingen uitgewisseld.	del centraal del extern del intern del sync and share opsl centraal opsl share and sync vind centraal
14:12	Interviewer: En als je iets groters moeten uitwisselen? Geïnterviewde: Wetransfer wordt veel gebruikt, maar veel groter dan dat zullen we niet uitwisselen.	del motivatie
14:14	Interviewer: En met wie wissel je uit? Alleen binnen Europa of met de Verenigde Staten.... Geïnterviewde: voornamelijk Europa, we doen veel Europese projecten binnen de Europese Unie. En dan zijn er tientallen partners, maar er zijn meestal maar een paar partners waarmee je echt samenwerkt op een bepaald onderwerp. Dat zijn er misschien vijf.	del extern del intern
14:15	Interviewer: Hoe komen die bij de data ook via dropbox achtige methodes? Geïnterviewde: Ja, of via de share schrijven op de case organisatie. Interviewer: Daar kunnen ze dan ook bij? Geïnterviewde: Daar kunnen ze bij.	del centraal del sync and share opsl centraal opsl share and sync
14:16	Interviewer: maar het is wel cruciaal denk ik dat het kan? Geïnterviewde: Dat het kan ja. Als voorbeeld. We zijn nog bezig met een paper te schrijven, verschillende codes van verschillende instituten, dus iedereen moet toch eigenlijk wel bij de data kunnen om ze te postprocessen. Dan gebruikt men die schijf. Interviewer: En spreken jullie dan ook bijvoorbeeld af dat ze alles bij jullie op de project schijf zetten? Geïnterviewde: Van wat ze dan willen uitwisselen wordt het [projectshare] dan als luik gebruik gewoon een handige methode	del centraal vind centraal

14:17	Interviewer: en heb je bijvoorbeeld naamconventies voor die bestanden of een folderstructuur? Geïnterviewde: Nee, dat laat ik aan de AIO's over. Ik weet dat je tegenwoordig van die naam conventies hebt binnen de data Management, hoe heet ze ook alweer? [naam van data steward]. In de praktijk doet men dat natuurlijk hoe ze dat willen.	fdst persoonlijk ind werkwijze indmot mate van flexibiliteit nmc persoonlijk
14:18	Interviewer: Metadata is de data over de data dus het kan zijn wanneer het geproduceerd is, op wat voor systemen het geproduceerd wordt. De naam van degene die het gemaakt heeft, de naam van het onderzoeksproject. Dat soort data zou je kunnen verzamelen en bij een dataset mee kunnen geven. Geïnterviewde: Nee, als ik er nu naar zou moeten kijken, als een AIO weg is dan archiveren we, dat gaat op naam. Dus is het wel duidelijk van wie het is. Dus dat hoeft niet, met wat het gegenereerd is. Wij zijn niet zo strikt in de zin van, Ik weet natuurlijk wel uit welke code het komt enzo, dat is meestal wel voor de hand liggend maar niet van welke precieze versie dat is geweest.	vlg code
14:19	Interviewer: En data documentatie dus echt een tekst document waarin staat beschreven wat de data allemaal inhoudt? Geïnterviewde: Dat komt er in de praktijk niet van. Interviewer: Ook niet in de vorm van comments in de code bijvoorbeeld? Geïnterviewde: Natuurlijk de codes zelf worden ook gearchiveerd. Dus die staan allemaal op bit bucket, onze bitbucket (een github achtig ding). Zo staan daar tientallen dingen een paar hele grote codes. Zelfs de papers worden daar gearchiveerd	archief ja archief waar ind werkwijze
14:20	Interviewer: en gebruik je dat ook voor versioning van zowel de code als van de papers? Geïnterviewde: Ja precies dus als er met meerdere mensen wordt gewerkt, wordt het via bucket gedaan en de data worden ook gelijk neergezet bij elkaar en weet je precies hoe het is gedaan als er een revisie moet komen. Kunnen we dat makkelijk doen	vlg code vlg ja vlg versie
14:21	Geïnterviewde: Omdat ik niet wilde dat....eh ik wilde geen open. Ik wilde die code niet open hebben. Dat is mijn competitive advantage. Dus ik wil dat niet open van mij. Interviewer: Wat bedoel je met competitive advantage? Geïnterviewde: Nou er zijn heel veel andere wetenschappers die vanalles doen en er zit heel veel tijd in die codes, dus die ga ik niet zomaar aan iedereen geven.	ind werkwijze
14:22	Interviewer: Maar is dit ook echt een wereld waarin data van elkaar gestolen wordt of dat men dat probeert? Geïnterviewde: Nee hoor, maar waarom zou je alles op straat gooien waar je jarenlang aan hebt gewerkt? Het is nog maar de vraag als je het wel doet, of mensen het dan gaan gebruiken want als relatieve leek, ja....eh... als relatieve leek, ... maar zelfs al ben je daar helemaal in, ga je het dan	ind werkwijze

	helemaal begrijpen wat er allemaal staat zonder..... dus in die zin hoef je misschien niet bang voor te zijn maar.	
14:23	Interviewer: Dus hoe zorg je dat alleen de mensen die bij de data kunnen komen waarvan jij wil dat ze er bij komen? Geïnterviewde: Nou ja die accounts op bitbucket. Die zijn alleen maar voor mensen die een account hebben en die geef ik alleen maar aan AIO's en studenten van mezelf. En daarna sluit ik ze ook weer. Dat is niet fullproof, elke student heeft daarbij gekund en heeft het dus in principe.	bvlg authenticatie bvlg autorisatie
14:24	Interviewer: Is het voor jou bijvoorbeeld ook belangrijk om achteraf te kunnen zien wie bepaalde bestanden benaderd heeft? Geïnterviewde: Nee, nee zo high security is het allemaal niet	logging
14:26	Interviewer: Maar nu zitten ze op bit bucket he? Hoe zit het met die wetenschappelijke data. Hoe zorg je dat daar alleen de mensen bij kunnen komen, bijvoorbeeld op zo'n project schijf? Geïnterviewde: Op zo'n project schijf heb je alleen toegang als je daar toegang toe hebt toch? Interviewer: En op HPC? Geïnterviewde: Op HPC kom je ook niet zomaar, daar moet je geactiveerde toegang toe hebben. Anders niet. Niet iedereen met netid heeft toegang.	bvlg authenticatie bvlg autorisatie
14:27	Geïnterviewde: Er zijn een aantal mensen die mogen vragen om toegang te laten activeren bij Robbert en Paul. Daar ben ik er ongetwijfeld één van, tenminste gaat altijd goed. Geïnterviewde: Dus ja je kunt niet zomaar vragen mag ik er op en rondkijken en dan werkt het niet.	bvlg authenticatie bvlg autorisatie
14:28	Interviewer: Het komt niet voor dat bijvoorbeeld een Amerikaanse wetenschapper of bijvoorbeeld iemand van buiten de EU waarmee je samenwerkt ook toegang tot hpc krijgt? Geïnterviewde: Heel af en toe maar in de praktijk zo ontzettend lastig dat dat niet zo vaak gebeurt, het is gewoon heel onhandig. Het systeem laat het niet makkelijk toe dat je het doet. Je krijgt niet makkelijk een netid namelijk je moet ergens al in peoplesoft staan. En als jij gast bent dan heb je daar niet zo makkelijk in dus dat komt niet zoveel voor.	bvlg autorisatie
14:29	Interviewer: De capaciteit om je data op te slaan. Hoe regel je dat je voldoende hebt? Geïnterviewde: Nou de hpc's van ons, de nodes zitten wel een beetje tegen hun limieten aan. Dat komt vooral omdat we niet zo snel archiveren waarschijnlijk, dus er staan nog wat oude spullen die eigenlijk weg moeten	opsl capaciteit opsl ext datahouder

14:30	Interviewer: En daar heb je wel faciliteiten voor, om die oude spullen weg te doen? Geïnterviewde: Nou die zouden naar de case organisatie schijven moeten gaan. Maar dan moet je dit wel weer gaan doen. Iedereen heeft het druk, komt er gewoon niet van. De ruimte opzich gaat volgens mij wel. Dat gaat	opsl capaciteit
14:31	Interviewer: En als je moeten uitbreiden? Misschien met alle respect voor het werk maar dat gaat dan meer AD HOC? Van ik heb meer schrijven nodig in mijn hpc? Geïnterviewde: Daar zou het op neerkomen, maar tot nu toe heeft het ons niet echt belemmerd in ons werk.	opsl capaciteit
14:32	Interviewer: En in bitbucket? Daar heb je gewoon voldoende ruimte? Geïnterviewde: Nou dat zijn meestal, meestal alleen de codes niet zozeer de datasets, voor de papers zijn de datasets heel beperkt. Het past allemaal wel.	opsl capaciteit opsl code
14:33	Interviewer: ook niet iets wat je nodig hebt of wenst? Geïnterviewde: misschien over een tijdje..... ik kan me voorstellen dat ze het misschien wel interessant vinden. Dat ze de steady state dataset voor de MSFR linken of een mesh dat ze denken van nu wil ik een transient doen en dan kan ik me het voorstellen, maar nu doen we dat niet.	dd link
14:34	Interviewer: Ik bedoel we hadden het net over codes die waarvan je niet wilt dat anderen die zien. Nou die hebben we dus ook. Die zijn dus op licentie van specifieke personen. Je tekent voor die code en die staan dus ook op het cluster. Ik zal niet zeggen dat dat nou helemaal netjes is zoals het moet zijn. Het is wel iets waar bij het nieuwe cluster misschien een beetje naar gekeken, moet worden dat er duidelijk groepen zijn die wel of niet....	bvlg authenticatie rd gevoelig
14:35	Interviewer: Dan staat er hooguit een agreement bij dat je er af moet blijven maar dat wil niet zeggen dat je de data niet kan openen? Geïnterviewde: Nou eigenlijk mag je het natuurlijk niet hebben. Je mag gewoon niet bij. Er zijn misschien niet zo netjes in. Maar tegelijkertijd is het ook een kwestie van wie heeft de tijd om dat te gaan regelen.	bvlg contract rd gevoelig
14:36	Interviewer: Ja wat ik nu in de paar interviews ook heb gemerkt is dat het veel van die samenwerking in de afdeling is gebaseerd op vertrouwen. Geïnterviewde: Ja tuurlijk weet je wat het ook is met sommige van die codes,daar moet je per persoon een licentie voor hebben. Dat betekent ook dat je per persoon een set discs moet aanvragen bij de NEA databank in Parijs of bij de OECD.	bvlg contract
14:37	Geïnterviewde: En als het onder licentie gaat dan betekent dat dat ik op mijn naam de code mag ophalen als ik heb aangetoond heb, dat ik ben wie ik ben en dat het ook echt aan mij wordt gegeven. Geïnterviewde: Ja, daar teken je voor. Bij wet heb je formeel een	bvlg contract

	groot probleem als je sommige van die codes distribueert. Nou laten we eerlijk zijn, iedereen heeft ze, dus iedereen die ze niet mag hebben ook, zo is het gewoon.	
14:40	Geïnterviewde: Het is altijd weer iets unieks Interviewer: het hoort ook bij onderzoek natuurlijk. Geïnterviewde: Dus het is niet zo van ik heb een weermodel ik ga de Oosterschelde van die en die dag doen of zoiets dat doen wij niet, maar altijd wel iets unieks altijd aan gesleuteld dus we hebben niet de standaard dataset van dan en dan.	indmot mate van flexibiliteit
14:41	Geïnterviewde: Maar mensen gebruiken allemaal hun eigen Dropbox.	ind werkwijze opsl share and sync
15:1	Geïnterviewde: Jazeker. We hadden het er eerder al even over dat het niet stopt als het onderzoek afgerond is, als een project klaar is. Dat daarna die data nog wel eens uit de kast wordt gehaald. Bij ons gebeurt dat niet zo heel vaak, dat we vragen krijgen na tien jaar van: Dat onderzoek dat jullie toen gedaan hebben. Daar willen we de data van zien.	archief eisen archief ja rd actief
15:2	Geïnterviewde: Wat wel nog eens gebeurt is dat delen van data uit een eerder project gebruikt worden voor het vervolg project of voor een project wat een afgeleide is van een eerder project. En dan blijft het dus wel actief. In die zin het gaat eigenlijk door, het transformeert naar een nieuw project.	rd actief rd herkomst
15:3	Geïnterviewde: De data komt van bedrijven, universiteiten en ziekenhuizen waar we mee samenwerken en daar zijn ook allemaal verschillende regels omheen.	rd herkomst
15:4	Geïnterviewde: Voor bedrijven geldt vaak een secracy agreement omdat dat concurrentie gevoelig is. Ziekenhuizen werken vaak met privacygevoelige data dus dat is een andere reden waarom je dan voorzichtig moet zijn ermee. En universiteiten dat is vaak wat makkelijker. Dat is vaak wat openbaarder	bvlg contract rd gevoelig
15:5	Geïnterviewde: Dat wordt, de ene keer wordt het meegeleverd door de PHD die langsgaat met een harddisk en die volstouwt met terabytes aan data. We hebben ook nog wel eens dat er cd's meegegeven worden. Komt wel eens voor bij ziekenhuizen. Steeds vaker komt het voor dat data in de cloud gedeeld wordt. Dat kan van alles zijn, we hebben alles wel zo'n beetje gezien. Zelfs ftp servers die dan ergens staan, waar je dan data vanaf kunt halen. USB sticks.	del datahouders del datatransfer del sync and share opsl ext datahouder
15:6	Geïnterviewde: En daarnaast, niet geheel onbelangrijk, we genereren onze eigen data in laboratoria met meetsysteem. Dat ligt heel erg aan de onderzoeksgroep en het soort onderzoek dat gedaan wordt. Dat kan terabytes per dag zijn, maar het kan ook een paar kilobyte per dag zijn. Daar zit alles zo'n beetje tussen.	rd herkomst rd meetdata

15:7	Interviewer: Het kunnen grote files zijn en kleine? Geïnterviewde: Ja kunnen ook gewoon een paar temperatuur meetwaardes zijn of een plaatje wat je genomen hebt die dag. Maar het kan ook zijn dat er continu, op hoge resolutie, met hoge snelheid imagedata moet worden opgenomen. Dat kan heel snel in de terabytes lopen, zeker als je daar lange metingen mee gaat doen. En dat is belangrijk omdat dat veel makkelijker is om te genereren en ook om te besluiten wat je daar vervolgens mee gaat doen.	opsl capaciteit rd meetdata
15:8	Geïnterviewde: Ook omdat wij een onderscheid maken in ieder geval in de groep CI maar ook bij MI, in data die goedkoop is, tussen aanhalingstekens en duur. Goedkope data definiëren wij als iets wat gewoon heel weinig kost of wat gratis is, maar wat je ook heel makkelijk opnieuw binnen kan halen of die je heel makkelijk kan meten, die maar een dag aan voorbereiding kost en een dag aan meettijd kost, tegen dure data waar je voor moet betalen, of die niet meer verkrijgbaar is omdat het van een patiënt komt waar we de data niet meer van kunnen hebben of die ook gewoon heel erg lang duurt om te genereren.	bvlg backup rd levels
15:9	Geïnterviewde: CI doet computational imaging en dat is ook voornamelijk computer werk. Maar, wat ik al zei, voor het binnenhalen van data moeten we soms wel onze eigen opstelling gebruiken. Dat gedeelte is puur experimenteel, want voordat je echt goede data hebt, ben je lang bezig om te werken aan je opstelling	rd herkomst rd meetdata rd simulatie
15:10	Geïnterviewde: Maar het verkrijgen van die data daar zijn we eigenlijk al jaren mee bezig. Dat loopt al twee jaar. En er komt ook data uit en er wordt al mee gewerkt, maar nog steeds zijn we bezig met het experimentele deel van het onderzoek.	rd meetdata
15:11	Geïnterviewde: maar het verkrijgen van de data, de transducers die daarvoor gebruikt worden, dat is nog steeds een groot belangrijk onderdeel van het onderzoek. Gisteren bijvoorbeeld was een jongen die doet borstonderzoek. Die is nu nog steeds bezig met de simulatie fase die gebruikt alleen nog maar computational, gegenereerde data, dus hij maakt zijn eigen samples maakt hij, die genereert hij gewoon aan de hand van onderzoek dat hij gedaan heeft. Dat kost overigens ook tijd, je bent lang bezig om een goed sample te kunnen maken, ook al is dat gesimuleerd. En pas na verloop van tijd zal de opstelling gereed zijn en dan kan die ook pas weten wat voor soort opstelling hij nodig heeft om dan experimentele data te gebruiken, daar komt ook weer een experimentatiefase aan te pas. Daarna komt er weer een computational fase. Dan moet hij gaan kijken of wat hij gedaan heeft, klopt met wat hij van tevoren al gesimuleerd had.	rd herkomst rd simulatie

15:12	Interviewer: En als we het hebben over experimental data. Een experiment suggereert, het kan goed of niet goed aflopen. Wat ik eerder heb gehoord is dat de experimenten die mislukken: Experiment 1 tot 25 mislukt en 26 lukt. Maar wat doe je dan met die datasets van die 25. Geïnterviewde: Probeer even een voorbeeld daarvan te bedenken. Daar heb ik eigenlijk te weinig inzicht in. Ik weet dat er weinig data wordt weggegooid. In ieder geval in deze twee groepen. Aan het eind van project wordt er wel een grote schifting gedaan en dan wordt ook inderdaad echt, de zeg maar zinloze data, de data die zeker na het project nooit meer gebruikt worden of niet door andere mensen, dat wordt echt wel weggegooid.	rd opruimen
15:13	Geïnterviewde: Daar moet je overigens wel, daar moet je actief actie op nemen. Je kan dat niet overlaten aan de zelfredzaamheid van een onderzoeker, dat moet je sturen, daar moet je opdracht toe geven. Maar dat gebeurt dan, terabytes worden weggegooid, 50, 100 terabyte wordt zo weggegooid.	ind werkwijze rd opruimen
15:14	Geïnterviewde: Tijdens het onderzoek is men voorzichtiger. Je bent toch bang dat er misschien wel zinvolle data tussen zit, ook al heb je dan je experimenten gedaan.	rd opruimen
15:15	Interviewer: En metadata in de actieve fase, doen jullie dat ook? Geïnterviewde: Steeds beter.	md ja
15:16	Geïnterviewde: Ja het besef dat het nodig is voor later voor het hergebruik van data. Maar ook het gebruik voor je zelf voor tijdens het schrijven van je verslag over je vak, vaak je proefschrift dat er gemaakt wordt. Ook omdat wij binnen onze groep meer en meer praten over het data plan dat we hebben. Een data management plan dat we nodig hebben	md definitie md ja
15:17	Geïnterviewde: Een data management plan dat we nodig hebben. Daar wordt al gesproken over, je moet kijken naar die metadata. En steeds vaker wordt er hier gewerkt met het neerzetten van je data op een archief. Dus bij een artikel publicatie moet je een DOI hebben. Dan moet je echt wel je metadata op orde hebben	archief eisen archief hoe archief ja md ja
15:18	Interviewer: Waar moet ik aan denken bij metadata bij jullie, wat versta je er onder? Geïnterviewde: Dat is heel verschillend omdat je de bronnen van de data zo verschillend zijn. Bij een experimentele data als over de Cryo gaat het over de temperatuur die op dat moment heerst, de spanningen die het op dat moment staan op bepaalde onderdelen van de opstelling, voltage.	md definitie md ja
15:19	Geïnterviewde: Ja. Maar bij patiënt data gaat het natuurlijk om de patiënten om de beschrijvingen van de patiënten die belangrijk zijn... Interviewer: Leeftijd, geslacht.. Geïnterviewde: Ja, dat soort dingen en eerdere aandoeningen, dus die zijn heel erg uiteenlopend	md definitie md ja

15:20	Interviewer: En datadocumentatie? Geïnterviewde: Nou, dat zie ik niet gebeuren. Interviewer: Nee? Geen readme of zo? Geïnterviewde: Nee dus het blijft bij metadata. De ene zet het in een schrift, een logboek en de ander zet het in een filteje en dat wordt dan wel readme file genoemd, maar daar staan alleen maar de gegevens in. Sommige zetten het in een Excel sheet, dan hebben ze een nummer gegeven aan een experiment en zetten daar ook bij wat de metadata is. Maar echte beschrijving, een tekstuele beschrijving dat zie je heel weinig.	dd readme ind werkwijze md definitie md ja
15:21	Geïnterviewde: We adviseren om de.....nou dit is een oud advies. We zijn wat aan het veranderen. Ik ga je eerst vertellen wat het oude advies was. Het oude advies was: Sla je data op op de groupshare en bulk storage. Sla de dure data op op de group storage, die is vaak ook kleiner. En sla de goedkope data op op de bulk storage.	opsl centraal
15:22	Geïnterviewde: Intussen het nieuwe beleid wat we willen en waar we mee bezig zijn om dat te introduceren is, we gaan werken project storage naar de umbrella. Sla daar je data op. Dat zal toch steeds voornamelijk de dure data zijn, want anders zitten we met problemen met de capaciteit van de storage, dus goedkope data zal nog steeds in de bulk staan, maar niet meer in de groupstorage.	opsl centraal
15:23	Geïnterviewde: Mensen die al twee of drie jaar bezig zijn. Dat is een beetje lastiger, die zitten al zo in het systeem. Die hebben al een externe harddisk hier en die hebben al een Dropbox, wat we al een paar jaar geleden besproken hebben.	opsl overal
15:25	Interviewer: En hoe worden die gedeeld dan? Met Git? Geïnterviewde: Nee nee nee nee dat wordt nog weinig gedaan. Ik probeer dat zelf wel nu te introduceren. Er zijn denk ik in de hele afdeling drie of vier projecten waarbij nu met Git ook scripts, voornamelijk scripts, worden gedeeld. Interviewer: Hoe delen ze het dan anders? Geïnterviewde: USB sticks of een printout van een stukje script dat geschreven is Interviewer: En geen Dropbox ofzo? Geïnterviewde: Mwoah ook wel, ja, nee dan wordt echt alles wel gebruikt, Dropbox, Surfdrive. Wat kom ik nog meer tegen? Met andere universiteiten maken we nog wel eens gebruik van SFTP. Nou ja, FTP dan, geen SFTP.	del centraal del datahouders del sync and share
15:26	Geïnterviewde: We hebben het nog niet over scripts gehad overigens. En dat is voor ons ook een heel belangrijk punt. Interviewer: Dat is ook onderzoeksdata of niet? Geïnterviewde: Dat is ehm, dat is onderzoeksdata, zeker onderzoeksdata en die wordt door alle trajecten die we besproken hebben, de simulatie fase, de experimentele fase, de meet fase en ook nog daarna de analyse die gedaan wordt, wordt bij ons gedaan op computers met computer programmatuur.	rd code rd definitie rd herkomst

15:27	Geïnterviewde: Dus we krijgen erg veel..... Aan het eind van het traject als het gaat over data die je dan wil archiveren. Daar zijn de scripts dan het belangrijkste van.	archief ja archief wat rd code rd definitie
15:28	Interviewer: En die scripts? Waar ze staan hangt misschien af van waarin je ze maakt maar die Python scripts staan in Git? Geïnterviewde: Meer en meer. Interviewer: En die Matlab scripts? Ook? Geïnterviewde: Nou wat minder. Matlab wordt toch echt gewoon, nou..... de meeste scripts worden volgens mij opgeslagen op de eh.... oh dat is interessant, dat zijn nieuwe ontwikkelingen op de groupstorage. Mensen die onder Windows werken, dat is echt een verschil, die werken dan op groupstorage. Dat zijn dan plekken die ze vanaf hun meetcomputers en hun desktop computer makkelijk kunnen benaderen. Maar omdat wij steeds vaker reken intensieve opdrachten krijgen waar onze desktop of workstations niet sterk genoeg voor zijn, werken we ook met onze HPC computers en die data die staat of op group storage, of op de local storage van de clusters en dat is verschillend voor Medical Imaging en voor Computational Imaging	opsl centraal opsl ext datahouder
15:29	Interviewer: Oké als je data deelt met wie deel je dan? Deel je alleen maar intern of ook extern, buiten de EU? Geïnterviewde: Even kijken, intern is het wat ik al zei vooral scripts of stukjes van scripts die gedeeld worden. Extern en dan met de partners waarmee samengewerkt wordt, data en scripts, ik scheid dat eventjes.	del extern del intern
15:30	Geïnterviewde: Scripts ook wel ik hoor ook van mensen die dan een script krijgen van een universiteit waar ze mee samenwerken. Een tijdje geleden moest iemand met het NKI scripts delen die ze nodig hadden. Voor de aansturing van de apparatuur worden soms scripts gedeeld. Dat gaat ook via alle wegen die er zijn. De ene keer is dat via Dropbox de andere keer is dat weer met een stickie dat wordt meegenomen.	del datahouders del sync and share opsl ext datahouder opsl share and sync
15:31	Interviewer: Maar ook met partijen buiten de EU? Geïnterviewde: Ook wel. Ik probeer even te verzinnen. Er is één heel bekend voorbeeld wat ik misschien al eerder genoemd heb. De groep computational Imaging heeft al een jaar of dertig een library ontwikkeld en die wordt wereldwijd beschikbaar gesteld en die wordt nog steeds opgehaald en gebruikt door alle landen die er zijn. Interviewer: En dat is een website? Geïnterviewde: We hebben een website voor en we gebruiken onze eigen FTP server die we binnen de case organisatie hebben neergezet.	del centraal del extern
15:32	Geïnterviewde: Voor de nieuwere versies daarvan gaan we gebruikmaken van GitHub. Daar is echt een complete rewrite van de software gemaakt, van de library. Die is van C overgezet naar C++ en	del code opsl code

	die zal niet meer op de case organisatie neergezet worden om op te halen die wordt via GitHub gedistribueerd.	
15:33	Interviewer: Hoe zorg je ervoor dat het veilig is? Geïnterviewde: Intern zorgen we ervoor met de permissies voor de Directories. Goeie vraag, extern zorgen we ervoor dat die data die gevoelig is, bijvoorbeeld patiënt data is een goed voorbeeld, dat hij niet beschikbaar is via kanalen die naar buiten gaan. Dus we hopen dan dat de beveiliging met permissies op de groupstorage voldoende zijn om de toegang van buitenaf daarmee te stellen	bvlg autorisatie
15:34	Interviewer: En encryptie gebruiken jullie dat? Heb je dat wel eens nodig? Geïnterviewde: Nee dat heb ik nog niet gehoord. Geïnterviewde: Er zijn er zijn twee promovendi die hebben op hun Apple computer filefault aangezet omdat zij wel willen dat hun data veilig is. Op die manier	bvlg encryptie
15:35	Interviewer: Zou het voor jou, jullie van belang zijn om te kunnen zien wie de data benaderd heeft, dat je dat achteraf kunt zien? Geïnterviewde: Incidenteel en daar kan ik wel een paar voorbeelden van noemen. Wij hebben bijvoorbeeld voor het ophalen van die library die we zelf gemaakt hebben. Vinden wij het fijn om te weten, in welke landen, voor welk platform het opgehaald wordt. En als we een nieuwe versie uitbrengen of het dan nog steeds wel levenswaardig is. Een ander voorbeeld is dat: we hebben, steeds vaker maken wij websites met presentaties van onderzoek wat bij een artikel wordt gepubliceerd zodat mensen bepaalde algoritmen kunnen uittesten via websites, daar maken we interactieve programma's voor. Dan vinden we het leuk om te zien, om te kijken, of we daar ook contacten mee kunnen binnenhalen. Is ook voor ons een soort visitekaartje wat je daarmee aflevert. En omdat je dan een log bijhoudt, kun je ook zien wie dan naar jouw kaartje gekeken heeft. En ten slotte omdat er wel wat vaker gebruik wordt gemaakt van Git via Github of Gitlab, heb je dat al gratis mee want mensen worden gelogd die op dat moment jouw spullen ophalen.	logging vlg share
15:36	Geïnterviewde: Wij gebruiken GitHub dat is extern en we gebruiken GitLab intern. Voorheen was het zo dat de GitLab omgeving op de case organisatie niet op public gezet kon worden. Tegenwoordig is het wel mogelijk met ICT'er voor elkaar gekregen. Zijn we bezig om de mensen die voorheen daarom op GitHub hun software zetten, nu ook naar GitLab te krijgen	del code opsl code

15:38	Interviewer: En naam conventies, folderstructuren? Geïnterviewde: Compleet arbitrair. Interviewer: Per onderzoeker bedoel je? Geïnterviewde: Per onderzoeker, nou [eigenlijk] per begeleider van onderzoeker. Dus het is wel zo dat een begeleider toch aangeeft wat zijn voorkeur heeft aan hoe de data opgeslagen wordt.	fdst groep fdst persoonlijk ind werkwijze nmc groep nmc persoonlijk
15:39	Geïnterviewde: Soms wordt daarbij ook aangegeven waar de data opgeslagen wordt in ieder geval de manier waarop het gestructureerd wordt. En dat is vaak op folder basis. De naambasis, ik weet niet of dat nou ook zo goed doorgevoerd is	fdst groep
15:40	Interviewer: En versies? Toen ik ooit hier aan begon, was het idee je begint met ruwe data dan doe je daar wat analyse overheen en dan is dat een beetje veranderd en heb je versie 2. daar doe je misschien nog manipulaties op en dan heb je versie 3. Is dat ook iets wat ook echt gebeurt en hou je dat dan bij? Geïnterviewde: Ik denk dat het iets, dit klinkt heel structureel. Ik denk dat het meer organisch is wat dat aan gaat. Het kan zo zijn dat het zo die structuur kan volgen dat het inderdaad een iteratie, bijna de waterval methode is. Op die manier. Dat is niet altijd met meetdata. Het is ook vaak zo dat een opstelling wordt gebruikt en er wordt een complete aanpassing gemaakt van de opstelling en ja dan heb je niet meer zoveel te maken met wat je daarvoor gemerkt hebt. Dat zou in versies toch niet zo veel zin hebben behalve dat zou kunnen duiden op een nieuwe opstelling.	ind werkwijze
15:41	Geïnterviewde: Het gaat mij er meer om dat je echt wat hebt gedaan met die data, je hebt er iets mee gedaan waardoor je eigenlijk een nieuwe dataset hebt gekregen. Je hebt er misschien dingen uit gefilterd? Geïnterviewde: Nee ik geloof niet dat het op die manier met versie wordt aangegeven	ind werkwijze
15:42	Interviewer: En datgene wat met datasets wordt gedaan, waar wordt dat bijgehouden? Houden mensen lab notebooks bij? Geïnterviewde: Ja, soms digitaal soms op schrift en soms niet. Dus alle drie. Digitaal, daar adviseren wij niet in. We hebben geen vaste manier waarop dat gedaan wordt dus iedereen doet dat op zijn eigen manier. Ik heb gezien dat mensen daarvoor notitieprogramma's gebruiken of inderdaad spreadsheets voor gebruiken. Sommige mensen een readmefile die bij de data staat waar dan beschreven staat wat er gedaan is met die data. En wat ik al zei: op schrift er worden logboeken bijgehouden. Wordt veel gedaan, trouwens. Veel logboeken, analoge logboeken.	dd data bewerken dd eln dd inhoud dd papier ind werkwijze vlg ja vlg transformatie

15:44	Interviewer: Versioning doe je wel voor scripts, maar dat is allemaal GitHub, GitLab. Geïnterviewde: Nee hoor, was het maar zo. Ik zou graag willen dat dat met Git veel meer gedaan krijgt met elke Version Control die je kan gebruiken. Subversion is ook nog steeds een hele goede En nee hoor ik heb van alles gezien. Daarbij, ik heb gezien dat een scriptfile met een underscore aan de filenaam met een v'tje en het nummer onder elkaar staan. Ik heb gezien dat het wordt gegroepeerd in verschillende folders met bijvoorbeeld de datum er in. Alle manieren die mensen maar kunnen verzinnen. En mensen die gewoon niet versiebeheer doen. Die dus ook gewoon hun data kwijt zijn.	vlg code vlg ja vlg versie
16:1	Geïnterviewde: Herkenbaar. Ik betwijfel of mensen het herkennen ons probleem.	opsl overal
16:2	Geïnterviewde: Ze zijn bezig met een onderzoek en denken vooral aan het publiceren van hun artikel, doen daarvoor testen op prototypes, er gebeuren metingen. Dat wordt gepubliceerd in een mooi tabelletje in het artikel. Alleen degene die het onderzoek uitvoert, als het een promovendus is, de promovendus en als het een postdoc is, de postdoc, weet waar die data is, waar de metingen staan	ind werkwijze rd levels
16:3	Geïnterviewde: het is heel erg persoonlijk die onderzoeksdata, er worden ook vaak een paar paden bewandeld. Proeven gedaan die voor niks zijn dat hoeft de onderzoeker niet altijd perse te bewaren. Totdat hij uiteindelijk de goede aanpak heeft gevonden die echt systematisch wordt doorgemeten, prototype goed getest	ind werkwijze
16:4	Geïnterviewde: Misschien met gebruikers bruikbaarheid, usability, misschien op performance getest, daar worden meetdata verzameld. En al die andere proeven daarvoor, ja dat zijn dan deelproeven afgebroken proeven waar wel wat metingen zijn die niet oninteressant zijn juist die alternatieven zijn wel aardig om achteraf te beschouwen.	rd herkomst rd meetdata
16:6	Geïnterviewde: Wij doen bijna niks anders dan met geografische data werken. Daar werken we continu mee maar die data zijn vaak data van overheden kadaster of Topografische Dienst of open source data, zoals streetmap maar dat zien we niet als ons onze data.	rd definitie rd herkomst
16:7	Geïnterviewde: Wij maken algoritmes en we gaan met die data aan de slag die bewerken we en we maken soms nieuwe data die we afleiden.	rd code rd herkomst rd levels

16:8	Geïnterviewde: het enige wat wij gedaan hebben is de data bij ons zo structureren dat die makkelijk via de webserver te distribueren is en in een browser makkelijk te bekijken is. Dus wat is onze data nou? Ja eigenlijk niks. We hebben de data gemasseerd gemanipuleerd in dat geval. Wat we als onze data nog zouden kunnen zien is hoe gebruiken mensen nu die AHN webviewer en welke queries doen ze , zoeken ze vaak in Delft, zoeken ze vaak In Utrecht, zoeken ze vaak in Maastricht, welke schaal zoeken ze, hoe vaak veranderen ze van schaal, dat is ook interessant voor ons. Het gedrag van gebruikers die kaarten gebruiken voor hun taken beter begrijpen.	del centraal rd definitie
16:9	Interviewer: Maar die samengevoegde data zie je dat niet als onderzoeksdata? Geïnterviewde: Dit is heel moeilijk een grens te trekken. Interviewer: Of de algoritmes waarmee je het samenvoegt? Geïnterviewde: Ja, die algoritmes, maar dat is software, dat zou ik geen data noemen dat moeten we ook bewaren en dat doen we meestal in GitHub of in een andere opensource eh...	rd code rd definitie
16:10	Interviewer: En scripts zijn in zekere zin ook onderzoeksdata zou ik zeggen vind jij van niet? Geïnterviewde: Ik vind het meer software. Een ander voorbeeld is wat we doen op variabele schaal onderzoek. We proberen van vaste kaartschalen af te gaan. Normaal heb je een schaal 1 op 10.000, 50.000 en 100.000 aparte kaartschalen. Dat zijn vaak 2D kaarten wat wij doen is proberen weg te gaan van die discrete vaste schalen maar een soort continu presentatie te maken, dat je op elke schaal de kaart kunt opvragen en de beste kaart voor die schaal te krijgen, bijvoorbeeld past goed op je beeldscherm. Goede informatiedichtheid. Wat we dan doen is beginnen met de meest nauwkeurige kaart schaal die we kunnen krijgen en daar gaan we algoritmes op loslaten en het resultaat stoppen we in een datastructuur dataset dat zou ik wel data noemen die voor al die schalen geschikt is. Voor een 2D kaart maken we een 3D datastructuur we zien de schalen als derde dimensie, dus een vlak wordt een volume en lijn objecten zoals wegen worden vlakken in de 3D ruimte. Verticaal is dan de Schaal, je moet dat zien alsof je papieren kaarten zou stapelen onderop leg je de meest gedetailleerde kaart daarboven de midden schaal en helemaal bovenaan op de kleinste schaal en daar kun je doorheen lopen met inzoomen en uitzoomen dat is eigenlijk een soort data kubus. Die maken we wel waar we beginnen met bron data En dan blijft het ook in dit geval moeilijk te zeggen: Dit is onze onderzoeksdata.	rd definitie
16:11	Geïnterviewde: Niemand anders heeft deze data maar een groot gedeelte van de data komt van een externe bron. Hetzij open source, open access, hetzij overheid.	rd herkomst

16:12	Geïnterviewde: Onze data is wat onze algoritmes toevoegen aan die oorspronkelijke data, is ook metingen die we doen met gebruikers van hoe goed ze taken uitvoeren, dat doen we ook wel op bescheiden schaal, Usability testen.	rd definitie
16:13	Geïnterviewde: En de grap is oké dat soort meetdata, dat zijn niet de enorme volumes. Misschien wel als je een webserver een jaar laat draaien en dan kijken welke ip adressen hebben welke queries gedaan. Dat begint dan wel aan te tikken in de loop der tijd. Maar als we met een groep van dertig personen ze per persoon een halfuur een taak geven van zoek dit of wat is de kortste route naar die locatie. Dan hebben we misschien maximaal een paar meg aan data, als we nog video recorden van die sessie wordt het ietsje meer.	rd herkomst
16:14	Geïnterviewde: Aan de ene kant die grote hoeveelheden die we gebruiken is wel onderzoeksdata of onderzoek heeft die data nodig en als we daar netjes data management afspreken wordt het ook erg duur dus het wordt meestal een compromis van de ene kant wel voldoende storage om te kunnen manipuleren en mee te experimenteren maar dat is geen archivering, lange termijn. Aan de andere kant een kleiner deel voor archivering voor die afgeleide meetdata, usability testen.	archief ja archief wat rd levels
16:16	Geïnterviewde: Ja wij hebben en een aantal servers hier in de groep. Eerst waren dat fysieke machines die we zelf hadden. En die staan in het rekencentrum	opsl ext datahouder
16:17	Geïnterviewde: Daarna zijn het meer virtuele machines geworden en daar hebben we nu een paar van.	opsl ext datahouder
16:18	Geïnterviewde: die gebruiken we ook wel een beetje voor productie achtige dingen als repository van onze eigen artikelen wat nu niet meer hoeft, Delft repository bestaat sinds 200 7, 8, 9	opsl centraal
16:19	Geïnterviewde: Nee gewoon Delft repository, waar alle scripties alle artikelen van TU Delft staan.	opsl centraal
16:20	Geïnterviewde: Toen wij begonnen met het opzetten van onze website bestond hij [Delft repository] nog niet, dat was 2000, en toen hebben we systematisch al onze artikelen gepubliceerd zet ook wel eens wat andere dingen neer zoals video's en die wordt ook goed geback-upped die machine, virtuele machine.	bvlg backup opsl centraal opsl ext datahouder
16:22	Geïnterviewde: We hebben nu elders geen data, we gebruiken niet het 4TU datacentrum. Ik heb het wel opgenomen in onderzoeksplannen van NWO maar dat is het vervolg op eerdere projecten met puntenwolken en varioschaal die tot nu toe nog niet goedgekeurd zijn	opsl ext datahouder
16:23	Geïnterviewde: Dus hebben we onze eigen ad hoc oplossingen gebruikt, onze eigen machines, onze eigen servers.	opsl ext datahouder

16:24	Interviewer: En iedereen in de groep heeft ook de discipline om het altijd maar daar neer te zetten niemand heeft meer wat op een laptop of op een externe harde schijf? Geïnterviewde: Nee dat is niet zo dat is niet zo. Elke persoon doet het weer op zijn eigen manier.	ind werkwijze
16:25	Geïnterviewde: anderen hebben de data of op een laptop, of desktop machine, of op een groepserver. We hebben ook nog wat netwerkschijven en die worden ook wel veel gebruikt. Dat moet ik eerlijk zeggen, dat zijn de Unix machines, die 4	opsl ext datahouder
16:26	Geïnterviewde: Daarnaast hebben de meeste mensen Windows, sommige hebben Mac desktop machines en sommige laptops. Daar hebben we een groepschijf, die heet bij ons M:, ik weet niet of dat bij alle faculteiten M: heet, maar bij ons heet hij M:. En daar staat ook het nodige, soms van de projecten staan er wel gegevens. Soms van vakken, bepaalde vakken waar men veel data moet delen met elkaar, dus dan staan daar gegevens.	opsl centraal opsl ext datahouder
16:27	Geïnterviewde: Uiteindelijk kun je zeggen is alles data, dus ik snap je opmerking van het begin van software of een script is dus ook data. Het zijn bytes. Aan de andere kant. Ik zie wel een groot verschil tussen data en software: Een algoritme. Eigenlijk zijn dat de twee ingrediënten: datastructuur en algoritmen dat maken samenwerkende systemen. Je kunt uiteindelijk, ook een algoritme worden bytes dus is ook data. De andere categorie, een publicatie, is ook data want uiteindelijk worden dat ook bytes die tekst die je typt en de plaatjes die er in staan is ook data maar ok is als je die categorieën meetelt dan is data management nog wat breder.	rd definitie
16:30	Interviewer: Ok. Dus het zou kunnen dat een promovendus bijvoorbeeld dat alleen maar op zijn laptop heeft staan. Geïnterviewde: Dat kan. Dat gebeurt ook.	ind werkwijze opsl ext datahouder
16:31	Geïnterviewde: We willen graag dat de software wordt hergebruikt en dan lopen we er ook tegen aan dat het lastig is als het zo persoonlijk is. Dat heeft twee gezichten: Aan de ene kant is dat gewoon de toegang, de access, weten waar het staat, het vindbaar zijn wat je ook al zei. Dat is één onderdeel, de andere kant is, het is gewoon moeilijk om de code van iemand anders te begrijpen en de data. Dat zal ook voor data gelden. Dan moet je er goede documentatie bij hebben.	bvlg autorisatie dd inhoud del motivatie
16:32	Geïnterviewde: En als je de moeite doet om de documentatie te lezen dan blijft het soms echt moeilijk om daar in te duiken. En het gevolg daarvan is, je ziet vaak dat mensen gewoon opnieuw beginnen en dat is zonde want je wil dat ze op de schouders van de collega's gaan staan of de voorgangers en verder gaan.	del datatransfer

16:33	Geïnterviewde: We hebben een sleutelonderzoeker, dat is [naam] op varioschaal gebied en alleen door hem lukt het dat dat overeind blijft, zal ik maar zeggen dat dat, het delen van de software en de data, het testen, de test datasetjes.	del motivatie
16:34	Interviewer: Want hoe doen jullie dat, dat delen? Geïnterviewde: Meestal gebeurt de code via GitHub, maar Martijn is weer een uitzondering, die gebruikt bit bucket.	del code opsl code
16:35	Geïnterviewde: Maar binnen onze groep is hij dus een uitzondering. Want studenten gebruiken vaak GitHub en Martijn doet bit bucket. Dus iedereen die met Martijn samenwerkt, doet dat op die manier. Interviewer: Zo deelt hij zijn code? Geïnterviewde: Ja zo deelt hij zijn code, ja	del code opsl code
16:36	Interviewer: En data delen? Geïnterviewde: Dat staat allemaal op de server, die machine. Ja daar zijn niet echt uitgeschreven spelregels voor maar dat is dus een directory. En daar kun je de datasets vinden en ja zelfs al zou je het niet snappen is het nog een soort logisch.	del centraal opsl centraal opsl ext datahouder
16:37	Geïnterviewde: Er zijn logische directorynamen gekozen Als het Duitse data is, staat er Germany, topographic data of pointclouddata. Dus voor het doel is het intern vindbaar en bruikbaar voor degene die hier op de TU zitten. Die machines zijn niet open van buiten, ze staan aardig dicht.	bvlg autorisatie fdst groep vind folderstructuur vind ja
16:38	Geïnterviewde: En ze gaan in Twente testen en kan die data makkelijk via de webserver worden gedeeld. Dat werkt wel goed. De gebruikersinterface, web Access, dat is wel een mooi platform om te delen.	del centraal
16:39	Geïnterviewde: De webserver is een geweldig platform. De toegang als die poort open staat op de webserver is wereldwijd. De webbrowsers met HTML5 en WebGL ondersteuning zijn zo extreem rijk aan functionaliteit en krachtig met 3D graphics en gebruiken de gpu erg goed. Zelfs als we iets voor intern bouwen, bouwen we meestal tegenwoordig een web oplossing en niet meer als een desktop specifieke oplossing	del centraal
16:40	Interviewer: Met het doel om te delen? Geïnterviewde: Met het doel het uiteindelijk makkelijk te kunnen delen maar het is.. De ontwikkelmoeite is nauwelijks meer. Ok wat je wel hebt je soms is de browser soms werkt het wel in Chrome en Firefox niet of omgekeerd	del centraal

16:41	Geïnterviewde: Dat was mijn zin die ik net begon. Wij willen in principe functionaliteit zo aanbieden dat mensen het gevoel hebben dat ze de applicatie via de webbrowser in handen hebben maar willen ze toch de data overhalen of om wat voor reden dan ook naar zich toetrekken. Weg bij de server bij ons, dan hadden we bij die puntenwolkdatasets een tool gebouwd dat je een rechthoek kon trekken.	del centraal
16:42	Geïnterviewde: omdat grote datasets kunnen zijn, kunnen die downloads ook flink duren. En dat kun je dan niet zo via e-mail sturen of zo, dus hadden we software gemaakt als je de rechthoek had getekend dan moest je zelf je e-mails adres opgeven, dan ging er bij ons een script aan het werk die die data ging verzamelen die stond vaak in onze structuur opgeslagen. Gingen we bij elkaar harken en in één groot bestand zetten, zetten we klaar bij ons op de server. En als hij klaar was. Nu is dat na een dag een mailtje. Staat jouw bestand klaar en je kunt het downloaden via deze ftp link, daarna is hij weer weg. Zo konden ze bij ons AHN data downloaden. Dan konden ze het niet alleen zien maar ook de data krijgen. Interviewer: Daar had je je eigen ftp server voor staan? Geïnterviewde: Daar hadden we onze eigen ftp server voor staan ja, een combinatie van http en ftp	del centraal del datatransfer
16:43	Geïnterviewde: …we er echt veel last van hebben dat elke keer dingen veranderen in browsers. Iets wat gisteren werkte hoeft vandaag niet meer te werken. Wij zijn een onderzoeksclub geen productieclub. Aan de ene kant is het leuk dat het gebruikt wordt. Aan de andere kant als ik met onderzoek bezig ben, zit ik niet te wachten op een telefoontje van ja, ik kan niet downloaden.	del motivatie
16:44	Geïnterviewde: Ik kreeg dus soms telefoontjes van de bibliotheek van ik heb hier studenten die kunnen niet downloaden. Dat vond ik dan leuk, kennelijk wordt het gebruikt. Dit is interessant onderzoek en zoveel data ontsluiten, echt terabytes en die data blijft trouwens maar komen want we hebben nu AHN twee en nu is AHN drie bijna af een paar jaar later. Het is eigenlijk nog leuker als je meerdere momenten in de tijd hebt want dan kun je zien wat er in Nederland is veranderd hoe de bomen gegroeid zijn en welke gebouwen er gesloopt zijn.	del motivatie
16:45	Interviewer: Maar daarnet had je het over de puntenwolk waarbij iedereen zijn stukjes kan bijdragen en daar kan ik me voorstellen dat dingen als Data Format en misschien wel metadata een hele belangrijke rol gaan spelen. Geïnterviewde: Ja standaardisatie is enorm belangrijk. Dat is cruciaal als je het formaat en ook als je het protocol van hoe je vraagt en antwoord geeft, niet standaardiseert dan werkt dat niet. In de basis hebben we alle W3C protocollen.	md ja md standaard

16:46	Interviewer: Wat voor protocollen zijn dat? Geïnterviewde: Van het World Wide Web Consortium, HTTP, FTP al het IP verkeer, mail adressen. Daarbovenop in onze wereld is het OGC, Open Geospatial Consortium belangrijk die heeft standaarden gemaakt voor de formaten maar ook voor de webservices. Geïnterviewde: Webmapservice, webfeatureservice, van data ophalen Geïnterviewde: Van metadata? Geïnterviewde: Die hebben ook metadata services, catalog services. Ik zou bijna zeggen dat dat die faciliteiten betreft, metadata, dat de geoinformatie wereld wel 10 20 jaar vooruitlopen op de rest ten aanzien van de informatie voorziening.	md standaard
16:48	Geïnterviewde: We hebben een geweldige service echt geweldig, het heet Pdoc publieke dienstverlening op de kaart is Nationaal Nederland echt het allerbeste wat je kunt voorstellen. Heel veel data te krijgen heel veel te vinden. Ja, werkt als een speer. Eigenlijk moet je twee dingen bekijken. Dat is Delft repository voor de publicaties van Delft en Pdoc dat is eigenlijk alle geografische data die in Nederland wordt gemaakt en die door overheden wordt gebruikt en daaruit wordt verspreid. Dus wij hoeven ons eigenlijk nooit zorgen te maken of weer een kopie van een topografische kaart moeten bewaren staat gewoon op Pdoc, kun je gewoon downloaden. Viewers, ze hebben ook mooie viewers gemaakt is echt heel goed gedaan. Echt heel goed gedaan.	rd herkomst
16:49	Geïnterviewde: Maar soms is vooruit lopen niet per se gunstig als er een grotere ontwikkeling komt met veel acceptatie. En dan moet je toch achteraf die ook ondersteunen.	md vorm
16:51	Geïnterviewde: Maar niet echt nagedacht over beveiliging en encryptie daarvan en het is altijd goed gegaan door de toegangssregels, wie er mag inloggen op de machine.	bvlg authenticatie bvlg autorisatie bvlg encryptie
16:52	Interviewer: En andersom? Wil je wel eens achteraf zien wie er aan de data heeft gezeten? Geïnterviewde: Ja. Bijvoorbeeld als ik echt de tijd zou hebben zou ik die logscripts van de http server willen zien. Ip adressen kun je zien.	logging
16:53	Geïnterviewde: Geautoriseerde toegang. Welke verzoeken, welke requests er zijn geweest welke geografische queries welke gebieden is er veel belangstelling voor. Is er nog niet van gekomen. Ik vind het wel jammer	logging
16:54	Geïnterviewde: Jouw vraag is zeer terecht, daar is veel belangstelling voor wie welke services en welke data gebruikt. Niet alleen bij ons maar ook in heel Nederland.	logging

16:56	Geïnterviewde: En naamconventies, folderstructuren? Daar heb je het net al eventjes over gehad, dat klonk heel persoonlijk vooral. Geïnterviewde: Ja, ad hoc gegroeid. Dus we hebben die M: schijf waar het aardig systematisch opgezet is met dingen met brieven namens de sectie, projecten waar meer mensen aan werken, binnen de sectie. Folder projectnaam en dan zie je een aantal projecten onder die folder. Education folder waar een aantal vakken onder vallen, een folder correspondentie waar de brieven onder vallen een folder notulen sectie overleg, waar de notulen onder vallen. Maar dat is niet geformaliseerd. Het is nu eenmaal zo.	fdst persoonlijk nmc persoonlijk vind folderstructuur vind ja
16:58	Interviewer: Maar voor je data en je software? Geïnterviewde: Heel veel staat op de Unix machines, op de Linux machines. Daar is het voor mij nu overzichtelijker, want ja in slash data, staat de data van de projecten. Maar als ik het niet zou weten en je zou mij bij de machine zetten. Ok, dan zou ik het wel achterhalen door een beetje te zoeken. Het zijn logische namen, maar dat is geen conventie gebruiken. Het is misschien heel anders gedaan op de AHN machines dan op de varioschaal machines.	nmc persoonlijk vind centraal vind ja vlg naamconventie
16:59	Interviewer: Maar jij hebt ook niet voor jezelf een eigen conventie, dat s per project anders? Geïnterviewde: Nee, behalve de gewone dingen zoals Temp is voor tijdelijk. Wat je ook vaak wel ziet is dat de data opgesplitst is in geografische gebieden en dat we coördinaten gebruiken om delen van data in verschillende directories te stoppen. Maar dat is eigenlijk wel per project anders, dat is de realiteit.	nmc persoonlijk vind folderstructuur vind ja vind naamconventie
16:60	Interviewer: En als je kijkt van het begin tot het eind wat er allemaal met die data gebeurt welke bewerkingen je toepast, welke algoritmes je er overheen gooit, gebruik je daar versiebeheer voor bijvoorbeeld? En zou je die levenswandel dan helemaal kunnen beschrijven? Geïnterviewde: Ja, we gebruiken zeker versie beheer. Maar versiebeheer op data durf ik niet te zeggen. Maar wel op de software op GitHub en Bit bucket. We hebben SVN draaien.	vlg code vlg ja
16:61	Interviewer: Maar voor data doe je dat niet zodat je bijvoorbeeld de eerste dataset een v1 geeft en na manipulatie een v2 en dat je tussentijds bijhoudt wat er is gebeurt? Geïnterviewde: Niet systematisch nee, er staat misschien, eh, als er al conventie is, is het dat er een source achter staat, dat het de brondata is, of de definitieve dataset die heeft dan de naam van die dataset wellicht. Final, maar dat durf ik niet eens met zekerheid te zeggen. Of dat we misschien een tussen versie 1 hebben. Wat ik zelf wel doe met al mijn documenten, maar dat vind ik dan minder data, ik gebruik extreem systematisch V1 en V2 V3 V4. Soms heb ik wel 100 versies staan. En dan denk ik: Ok is wel een beetje veel.	ind werkwijze vlg naamconventie vlg versie

16:62	Geïnterviewde: We hebben wel een project en dat is wel interessant trouwens. Dat is minder, is het data, is het geen data? Werkt als standaardisatie, OGC is de standaardisatiemix van de industrie, de overheid en de onderzoekswereld in de geo informatiewereld. Je hebt de ISO voor wereld standaarden. Ik werk met ISO mee aan 1 standaard, die heet Land Administration Domain Model en we zijn net aan een revisie bezig.	md standaard
16:63	Geïnterviewde: Die stelt dit voor, die stelt dit voor en dan heb je twee versies van het model. En daar hebben we ook eigenlijk die lompe aanpak van ons met v1, v2, v3, v4 en 1 gedeelde directory waar we nu wel met met het probleem zitten, dat mensen van buitenaf er niet makkelijk bij kunnen. Die kunnen niet makkelijk bij onze M: schijf. Dan moeten we het toch opsturen op een andere manier.	nmc persoonlijk vind ja vind naamconventie vlg versie
16:64	Geïnterviewde: Oké ik had eerst ook nog, en dat doe ik nog steeds vrij veel: de Wiki van de case organisatie, gebruik ik ook vrij veel. Maar daar zit ik steeds met 10 Meg als grens van een file. Dat was ook de reden dat we van de wiki zijn afgestapt voor de LADM modellen. Op een gegeven moment waren die 20Meg, 30Meg, 40Meg.	del centraal opsl centraal
16:65	Interviewer: Gebruik je wel eens Surfdrive of Dropbox om dingen te delen? Geïnterviewde: Gebeurt ook, wat dat betreft gebeurt alles.	del sync and share opsl share and sync
16:66	Interviewer: Maar jij niet? Geïnterviewde: Ik gebruik het zelf als iemand mij het opstuurt. Ook omdat hier bij de TU zelf zulke goede faciliteiten hebben. Als ik zelf een grote file heb, bijvoorbeeld ik scan een scriptie van iemand met mijn commentaar erop. Dat zou je anders misschien via Dropbox doen of WeTransfer. Maar ja, ik heb mijn eigen server hier, ik zet het gewoon in mijn homedirectory, die publiek is, althans in dat deel dat publiek is en ik stuur de link door naar de student.	del sync and share
16:67	Geïnterviewde: Maar als ik niet die goede faciliteiten van de TU zou hebben, zou ik meer Dropbox of Wetransfer gebruiken waarschijnlijk.	del sync and share

16:68	Geïnterviewde: Gedurende het gesprek zijn me steeds meer dingen te binnen geschoten. Die ik niet voorzien had. Wat ik zie is dat er wel een tsunami is aan data. Dat wat we tot nu toe dachten dat veel was. Dat is over vijf jaar echt heel weinig data. Dus het promotieonderzoek van X die met die puntenwolken bezig is, hebben we gezegd, in de toekomst verwachten we dat we tien tot vijftiende punten moeten beheren. Tien tot de negende is een miljard, 10 tot 12de is een triljard. En AHN nu is bijna 1 triljard. We denken dat het nog wel duizend keer zoveel zal worden, wat een organisatie wellicht voor een land moet gaan beheren. Dat is maar één type data. Maar die puntenwolken zijn wel heel veel data. En dat zie je ook met sensoren. Sensordata, bewaar je dat wel of niet? Wat we ook als geodata zien zijn sensoren die her en der geïnstalleerd zijn op vaste plekken of die in bewegende objecten zitten in auto's en die ook data verzamelen met een locatie erbij en die kun je ook delen. Er zijn ook standaarden voor maar die worden niet altijd gebruikt. Bijvoorbeeld OGC standaarden, Ja die sensordata, ontiegelijk nuttig en rijk als je die goed kunt gebruiken. Maar het is een enorme hoeveelheid data.	opsl capaciteit
16:69	Geïnterviewde: En dan ook d de vraag is dat onderzoeksdata? Ja wel, als je die wilt analyseren en met machine learning patronen wilt ontdekken. Wat zie ik in die sensordata? Er zijn voorbeelden zoals sensoren geluid of luchtvervuiling meten soms doen burgers dat zelf, maar dat kan best wel privacygevoelig zijn niet. Dat je denkt het is niet zo privacygevoelig, maar uiteindelijk is het eigenlijk wel. Het sensornetwerk wordt zo dicht dat je kan zien dat iemand ongeveer op dat adres, elke dag om dat tijdstip, met zijn motor vertrekt en dan die kant oprijdt, met geluidssensoren, als je die verbindt in een netwerk, met machine learning haal je dat eruit. De beveiliging en privacy dingen waarvan we zeggen: Ok, daar doen we nu niet al te veel aan. Dat wordt dan ook wel een vraagstuk.	rd definitie rd gevoelig